NAME:ANUSHKA SUR

EMAIL:anushkasur35@gmail.com

Bank Loan Case Study

Project Description

This case study aims to give you an idea of applying EDA in a real business scenario. The problem at hand is to analyse a dataset containing information about clients and their loan applications. The objective is to identify factors that differentiate clients with payment difficulties from other cases. By conducting a detailed analysis, we aim to gain insights into the characteristics and patterns associated with clients who face payment difficulties, which can help in developing strategies for risk assessment and decision-making in the lending process.

Approach used:-

The analysis is divided into three main steps:

1. Exploratory Analysis and Data Cleaning:
    - The first step involves exploring the dataset to gain a better understanding of its structure and contents.
    - During this phase, data cleaning techniques are applied to handle missing values, outliers, and inconsistencies.
    - Key columns that are relevant for the analysis are identified, based on their importance and relationship to the target variable.
2. Univariate, Bivariate, and Multivariate Analysis:
    - In this step, various types of analysis are performed on both categorical and numerical variables.
    - Univariate analysis focuses on understanding the distribution, central tendencies, and variability of individual variables.
    - Bivariate analysis examines the relationship between the target variable and other variables to identify any patterns or correlations.
3. Identifying Predictive Variables for High-Risk Customers:

4. Based on domain knowledge and insights gained from previous analyses, specific variables that have the potential to predict high-risk customers are identified.
5. Variables such as 'TARGET' (indicating delayed payments) and 'NAME_CONTRACT_STATUS' (previous loan application status) are mentioned as important variables for analysis.
6. These variables will be further analyzed to determine their predictive power and contribution to identifying high-risk customers.
7. The purpose of this analysis is to understand the dataset, clean the data, explore relationships between variables, and identify key variables that can help predict high-risk customers. By following these steps, meaningful insights can be derived to inform decision-making and risk assessment in relation to loan applicants

8. List the top ten correlations between TARGET variables.All of the meaningful data (analysis work) was finally presented in the form of various graphs and chart

Tech-Stack Used

Google Colab:- Google Colaboratory ("Colab" for short) is a data analysis and machine learning tool that allows you to combine executable Python code and rich text along with charts, images, HTML, LaTeX, and more into a single document stored in Google Drive. Python (Programming Language):- Python is a programming language that has extensive support of libraries which makes data analysis easier.

Answers

Understanding Data:

1. `application_data.csv` contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

2. `previous_application.csv` contains information about the client's previous loan data.

It includes the data on whether the previous application had been Approved, Cancelled, Refused, or Unused offer.

3. `columns_descrption.csv` is a data dictionary that describes the meaning of the variables.

Identify the missing data and use appropriate methods to deal with it. (Remove columns/or replace them with an appropriate value)

There are two datasets namely ‘Applicaton_Data’ and ‘Previous_Application’.

2. Initially calculated the null value percentage of each column and found out the columns

containing null values above 40% in Application data are 64 columns and in Previous Application are 11 columns.

In Application Data –

We eliminated columns from the dataset that had missing values greater than 40%.

In the "application_data" dataset, we identified the "AMT_GOODS_PRICE" column as valuable for data analysis.

To handle the missing values in this column, we used the median operation to impute them.

Similarly, we also imputed the missing values in the "AMT_ANNUITY" column using the median operation.

In median imputation, the missing values are replaced with the median value of the entire column.

Additionally, we removed columns such as 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_DOCUMENT_1', and others that were deemed irrelevant for the analysis work from the data frame.

In Previous application.

We adjusted the columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', and 'DAYS_ID_PUBLISH' to convert negative values to positive. This was done because days cannot have negative values.

In the "Previous Application" dataset, we observed that columns such as 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'AMT_DOWN_PAYMENT', and 'CNT_PAYMENT' had null or missing values. To impute these null values, we used different methods:

- For the 'AMT_ANNUITY' column, we filled the null values using the median operation, which means replacing the missing values with the median value of the column.

- In the case of the 'AMT_GOODS_PRICE' and 'AMT_CREDIT' columns, we used the mode operation. Mode imputation involves filling the missing values with the most frequent value in the respective columns.
- For the 'CNT_PAYMENT' column, we filled the null values with '0', implying that missing values were replaced with zero.

These imputation techniques were applied to address the missing values and ensure that the dataset is more complete and suitable for analysis.

Identify if there are outliers **in the dataset. Also, mention why do you think it is an outlier.**

Outliers are data points that deviate significantly from other observations in a random sample from a population. In this analysis, we identify outliers using box plots, which provide a visual representation of statistical data including the minimum, first quartile, median, third quartile, and maximum values.

A box plot displays these key statistical measures as a box, with the bottom edge representing the first quartile (25th percentile) and the top edge representing the third quartile (75th percentile). The median is depicted as a line inside the box. Additionally, the plot includes whiskers that extend to the minimum and maximum values within a certain range.

Outliers, which are observations that lie outside this range, are displayed as individual points outside the plot, symbolizing their abnormal distance from the rest of the data. These points represent data points that are significantly different from the majority of the observations and may require further investigation or special consideration in the analysis.

By using box plots to identify outliers, we can visually assess the distribution of the data, locate extreme values, and gain insights into

any potential anomalies or data points that might have a significant impact on the overall analysis.

APPLICATION_DATA.CSV

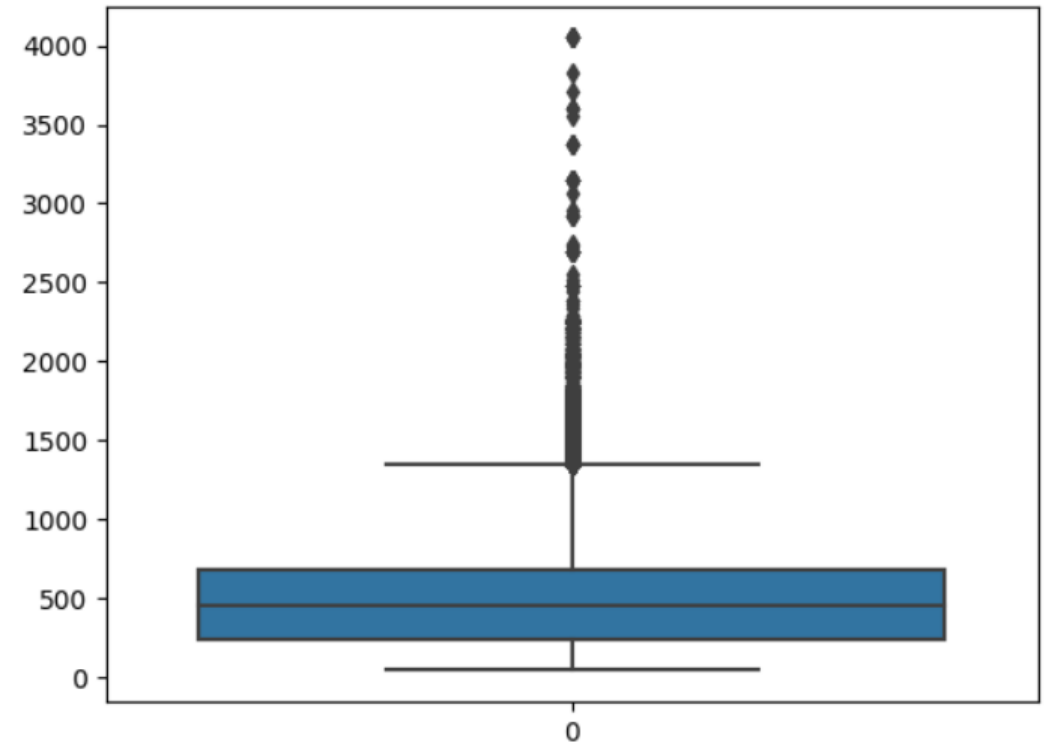
OUTLIERS:-

For the outlier analysis of numerical columns, we will focus on

- AMT_GOODS_PRICE

```
# dividing by 1000 for the ease of read and converting value in ('000s')
sns.boxplot(app['AMT_GOODS_PRICE']/1000.0)
```

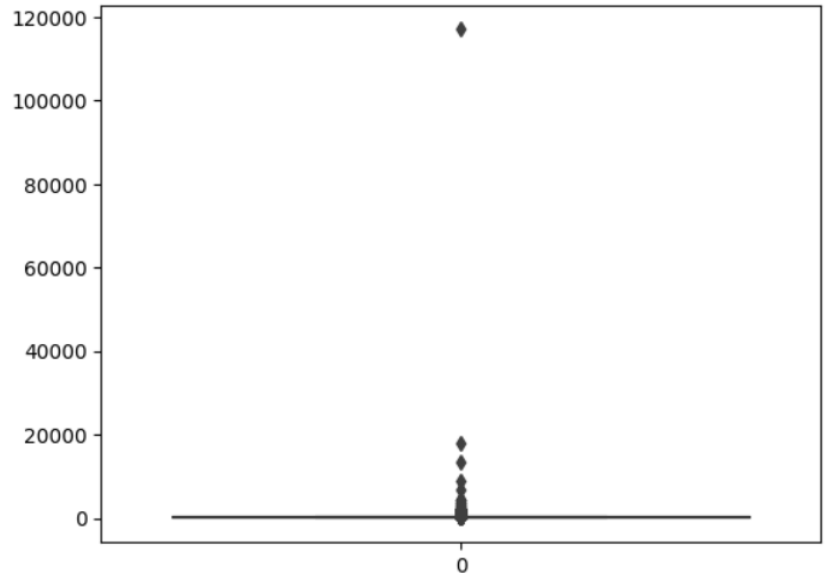
<Axes: >



- AMT_INCOME_TOTAL

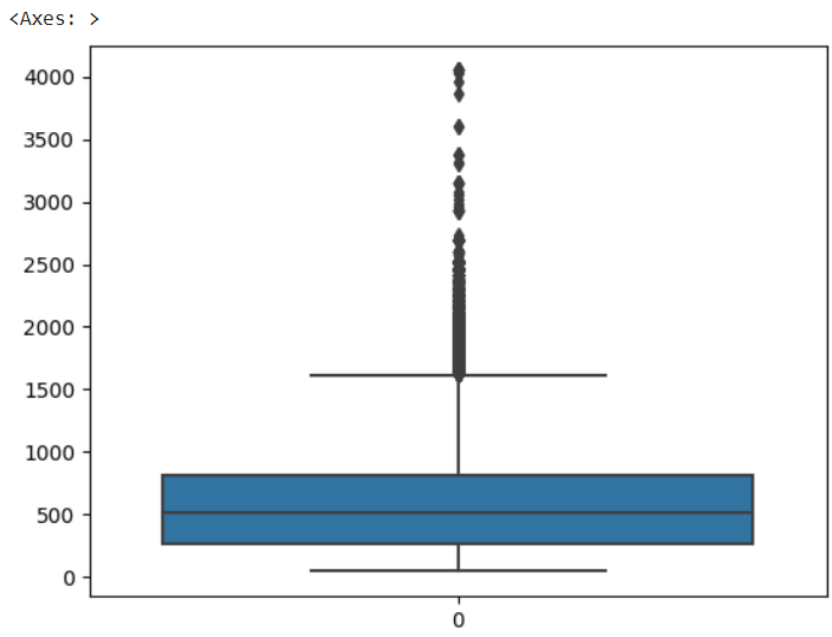
```
# dividing by 1000 for the ease of read and converting value in ('000s')
sns.boxplot(app['AMT_INCOME_TOTAL']/1000)
```

```
<Axes: >
```

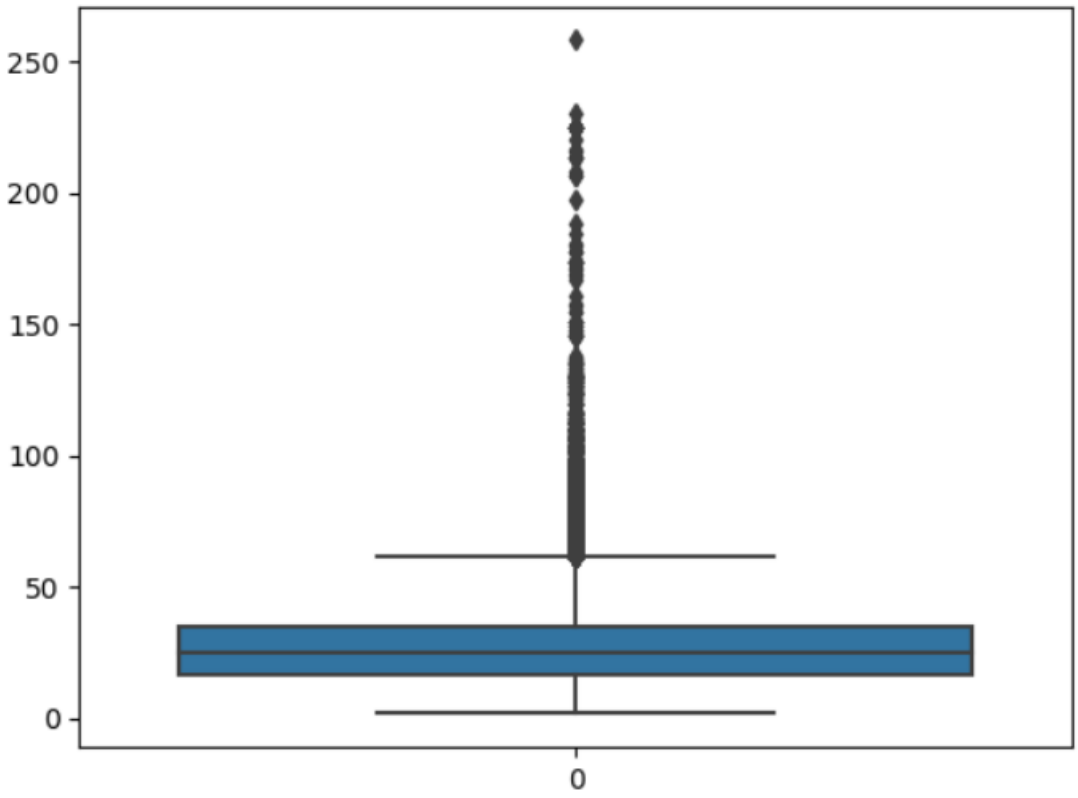


- AMT_CREDIT

```
<Axes: >
```

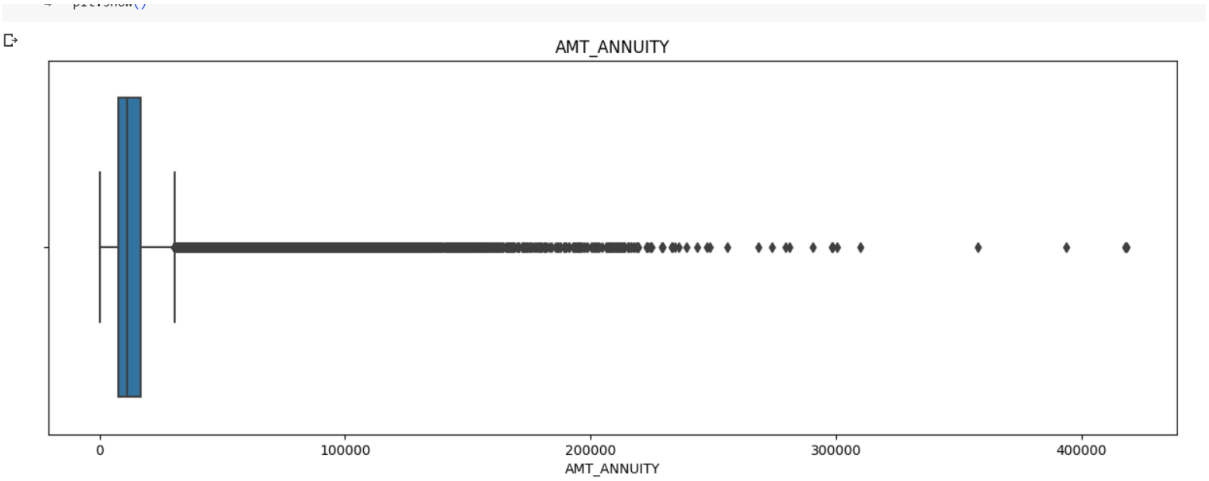


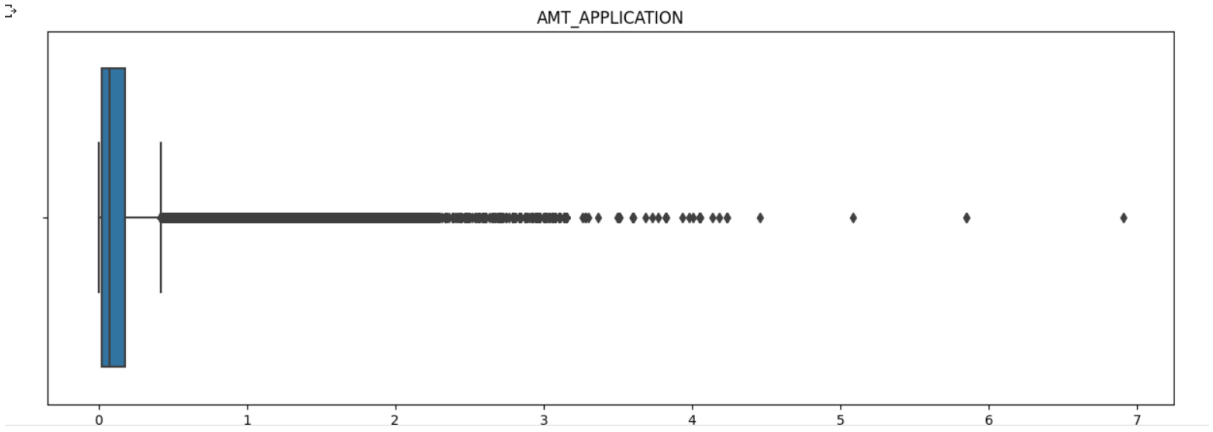
- AMT_ANNUITY

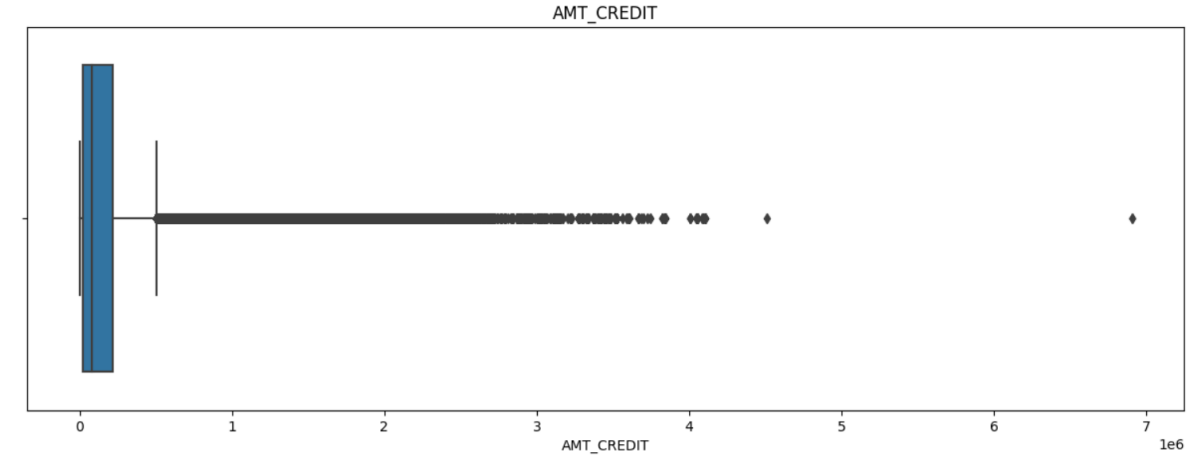


IN PREVIOUS_DATA.CSV:-

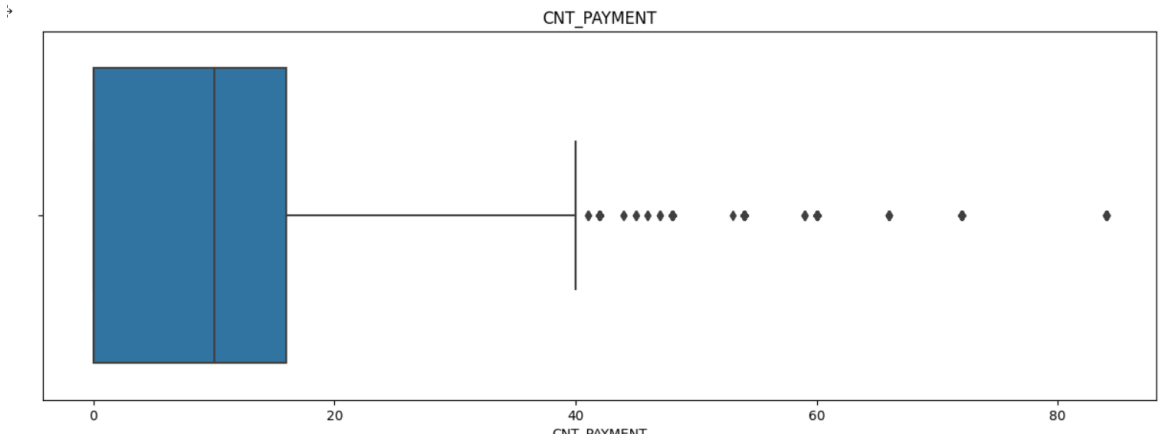
OUTLIERS:-

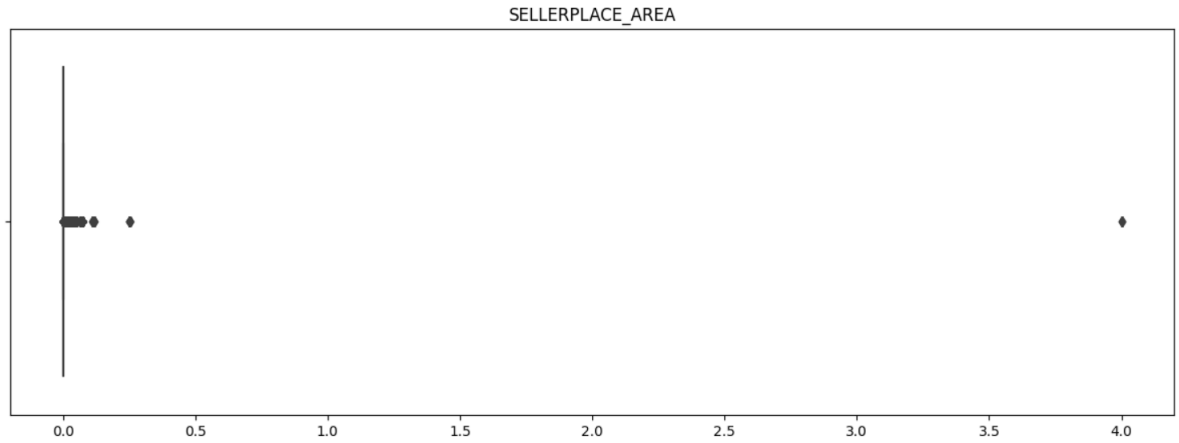



AMT_APPLICATION
0
1
2
3
4
5
6
7


AMT_CREDIT
0
1
2
3
4
5
6
7
AMT_CREDIT
1e6


AMT_GOODS_PRICE
0
1
2
3
4
5
6
7
AMT_GOODS_PRICE
1e6





ANALYSIS:-

Application data

1. AMT_ANNUITY, AMT_CREDIT, have some number of outliers: The presence of outliers in these variables suggests that there are loan applicants who deviate significantly from the majority of the applicants in terms of annuity amount, credit amount, or the number of children.
2. These outliers can have a notable impact on statistical analyses and should be carefully considered. It's important to investigate these outliers further to determine whether they are legitimate data points or potential errors

.Previous data

3. AMT_INCOME_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high incomes compared to the others in the case of application data.
4. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA have a huge number of outliers.
5. . CNT_PAYMENT has few outlier values.

Identify if there is a data imbalance in the data. Find the ratio of data imbalance.

The result of data imbalance shows The ratio of data imbalance relative with respect to Repayor and Defaulter data is 11.39

As the percentage of Target =0 and Target =1 are different, there is an imbalance

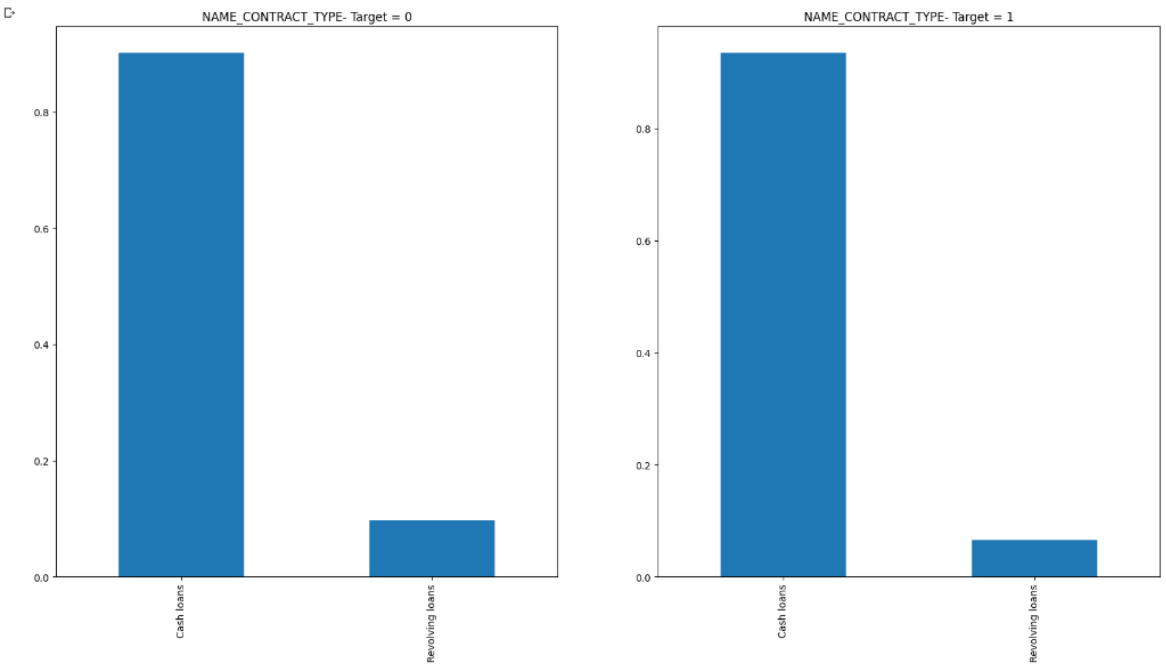
Explain the results of univariate, segmented univariate, bivariate analysis, etc.in business terms.

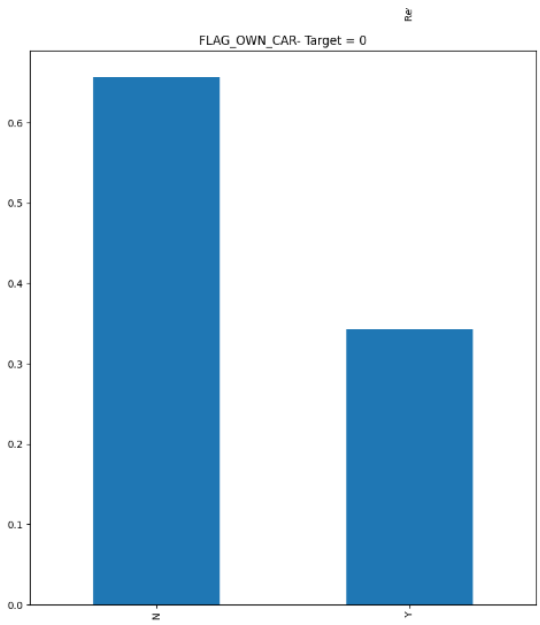
Any column which is either of object type or have less than 40 values been considered categorical. Remaining columns of type float or int will be considered numerical

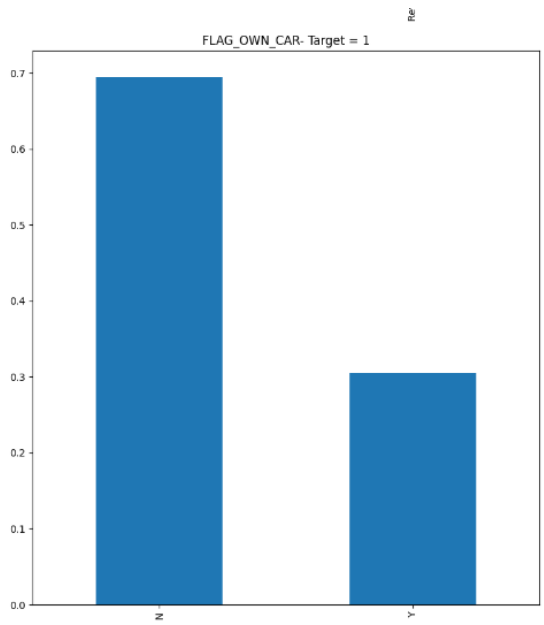
Univariate Analysis for categorical variable

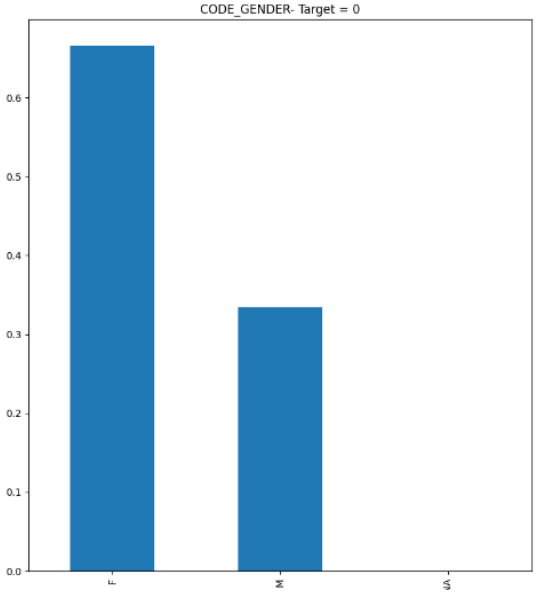
```
list of all categorical columns
categorical_columns = ['NAME_CONTRACT_TYPE',
'FLAG_OWN_CAR',
```

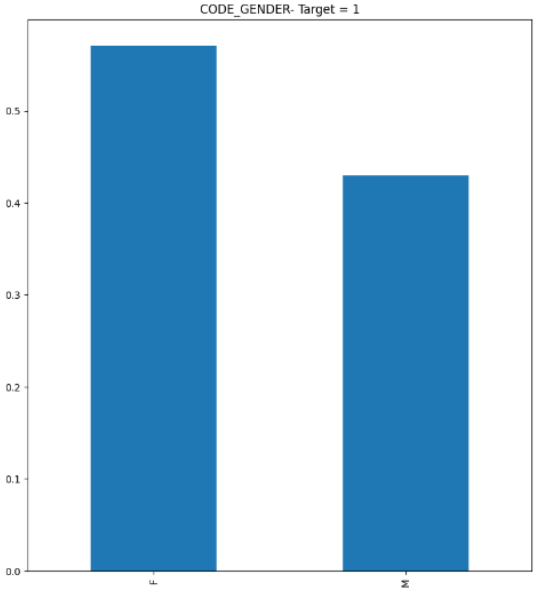
```
'FLAG_OWN_REALTY',
'CODE_GENDER',
'NAME_EDUCATION_TYPE',
'AMT_CATEGORY',
'AGE_GROUP',
'NAME_FAMILY_STATUS',
'NAME_HOUSING_TYPE',
'NAME_TYPE_SUITE',
'NAME_INCOME_TYPE',
'OCCUPATION_TYPE',
'ORGANIZATION_TYPE',
'REGION_RATING_CLIENT_W_CITY',
'REGION_RATING_CLIENT',
'AMT_REQ_CREDIT_BUREAU_HOUR',
'AMT_REQ_CREDIT_BUREAU_WEEK',
'AMT_REQ_CREDIT_BUREAU_DAY',
'DEF_30_CNT_SOCIAL_CIRCLE',
'AMT_REQ_CREDIT_BUREAU_QRT',
'CNT_CHILDREN',
'CNT_FAM_MEMBERS',
'AMT_REQ_CREDIT_BUREAU_MON',
'AMT_REQ_CREDIT_BUREAU_YEAR',

]
```

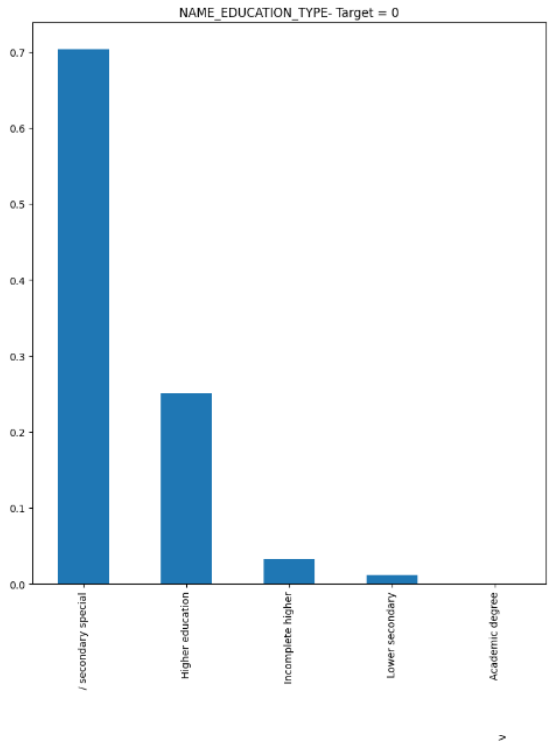


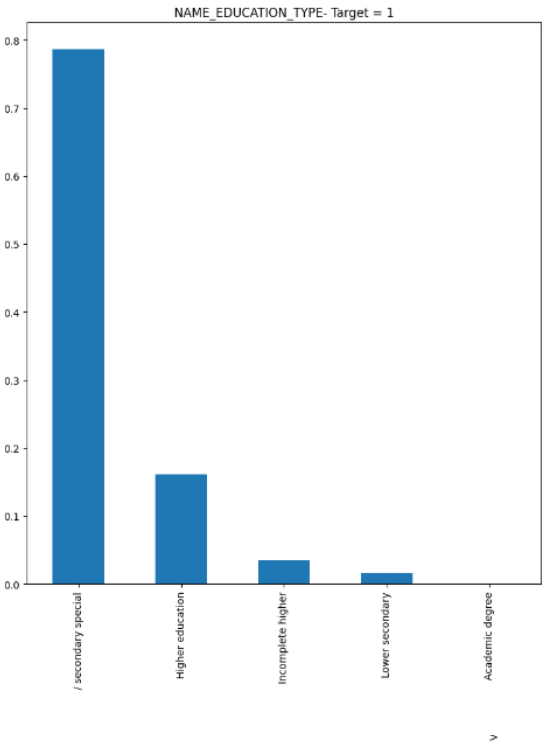

Re
FLAG_OWN_CAR- Target = 0
0.6
0.5
0.4
0.3
0.2
0.1
0.0
N
Y

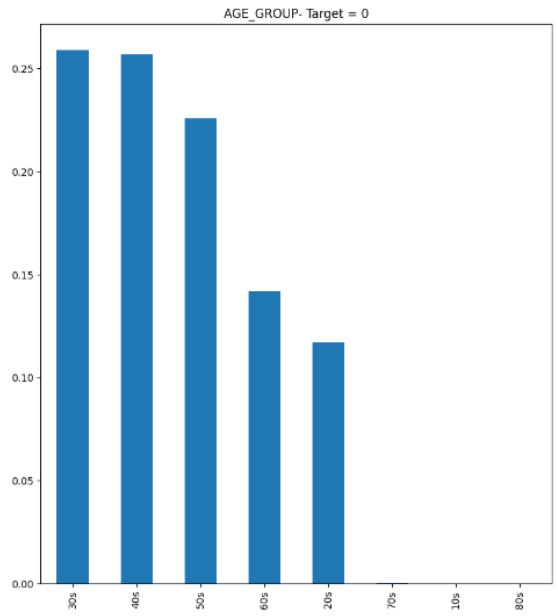

Re
FLAG_OWN_CAR- Target = 1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
N
Y

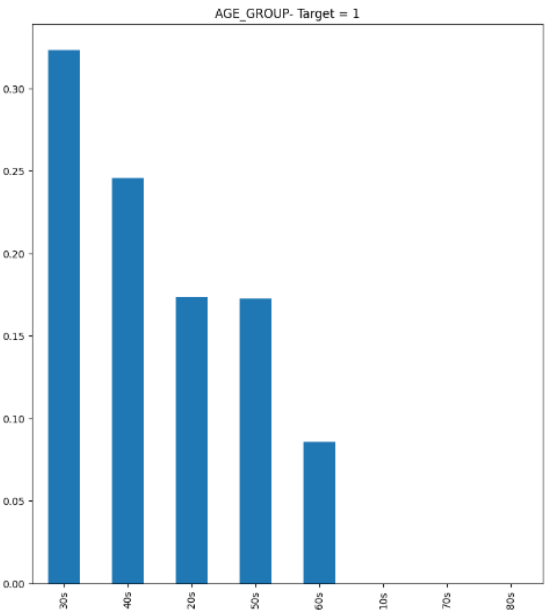

CODE_GENDER- Target = 0
0.6
0.5
0.4
0.3
0.2
0.1
0.0
F
M
XA

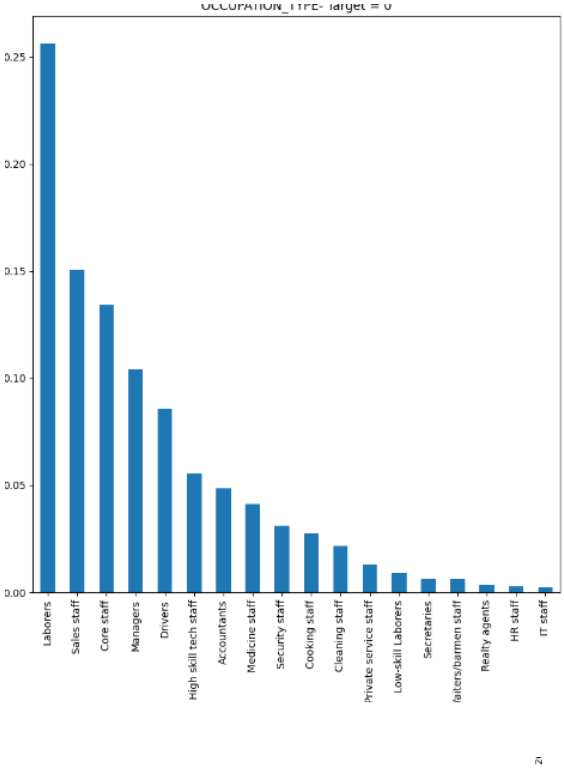

CODE_GENDER- Target = 1
0.5
0.4
0.3
0.2
0.1
0.0
F
M

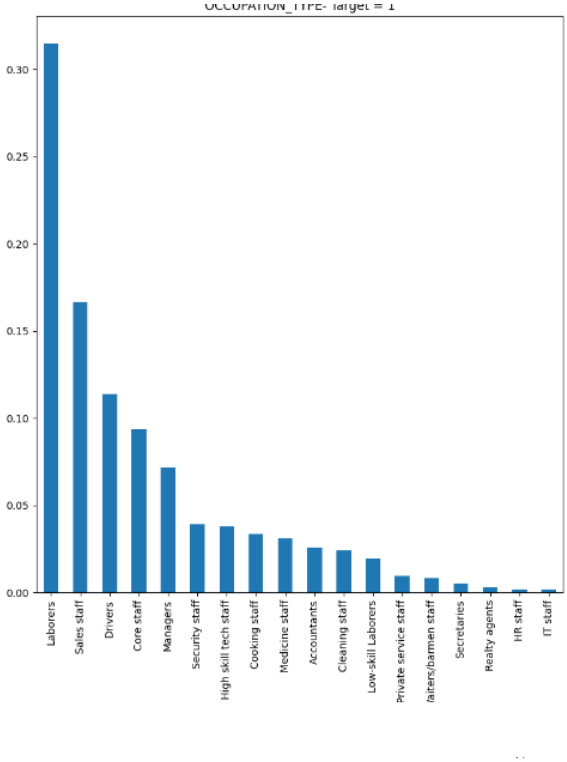

NAME_EDUCATION_TYPE- Target = 0
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
/ secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

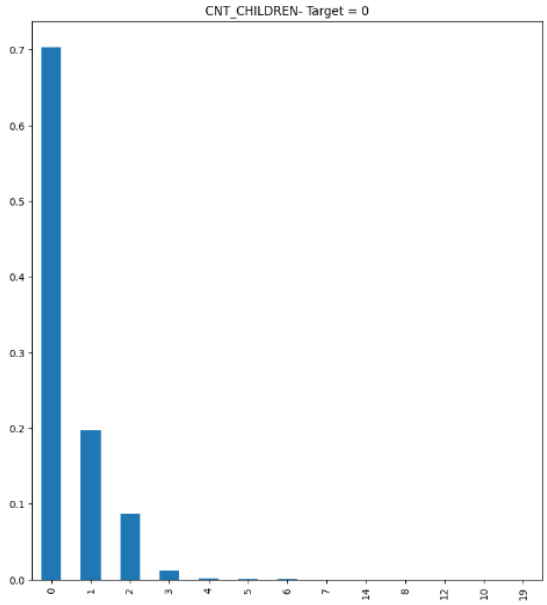

NAME_EDUCATION_TYPE- Target = 1
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
/ secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

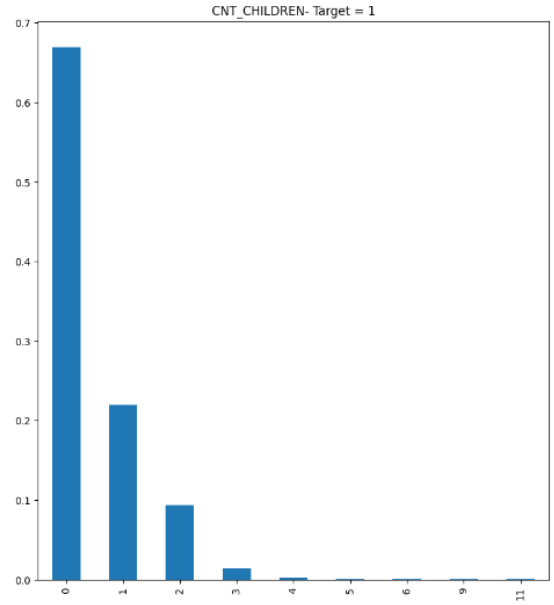

AGE_GROUP- Target = 0
0.25
0.20
0.15
0.10
0.05
0.00
30s
40s
50s
60s
20s
70s
10s
80s

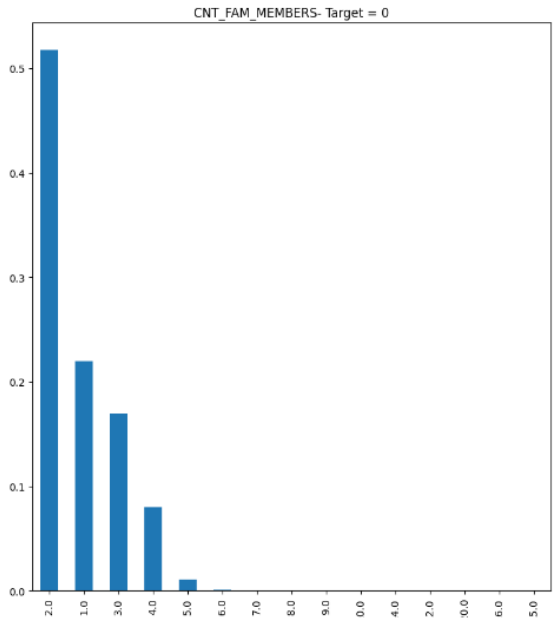

AGE_GROUP- Target = 1
0.30
0.25
0.20
0.15
0.10
0.05
0.00
30s
40s
20s
50s
60s
10s
70s
80s

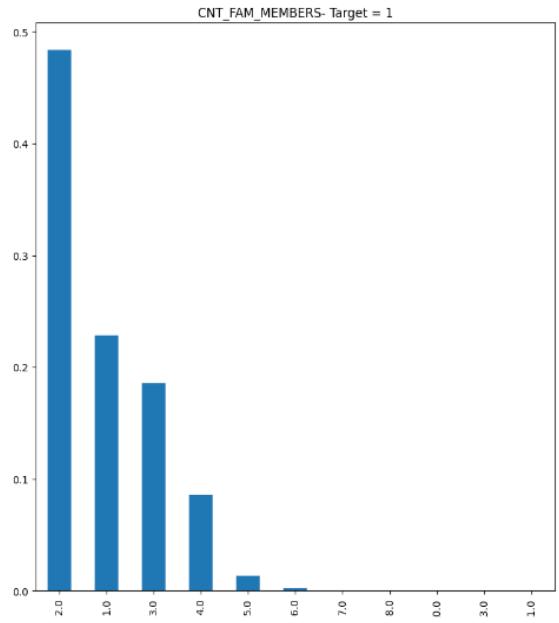

OCCUPATION_TYPE- Target = 0
0.25
0.20
0.15
0.10
0.05
0.00
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Secretaries
Waiters/barmen staff
Realty agents
HR staff
IT staff


OCCUPATION_TYPE- Target = 1
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Laborers
Sales staff
Drivers
Core staff
Managers
Security staff
High skill tech staff
Cooking staff
Medicine staff
Accountants
Cleaning staff
Low-skill Laborers
Private service staff
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff


CNT_CHILDREN- Target = 0
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
14
8
12
10
19


CNT_CHILDREN- Target = 1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
9
11



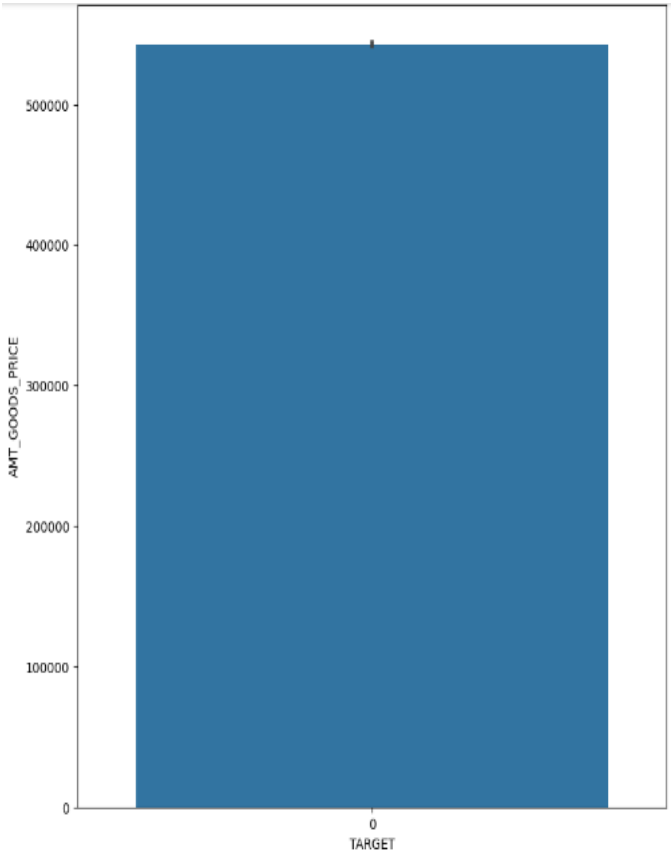


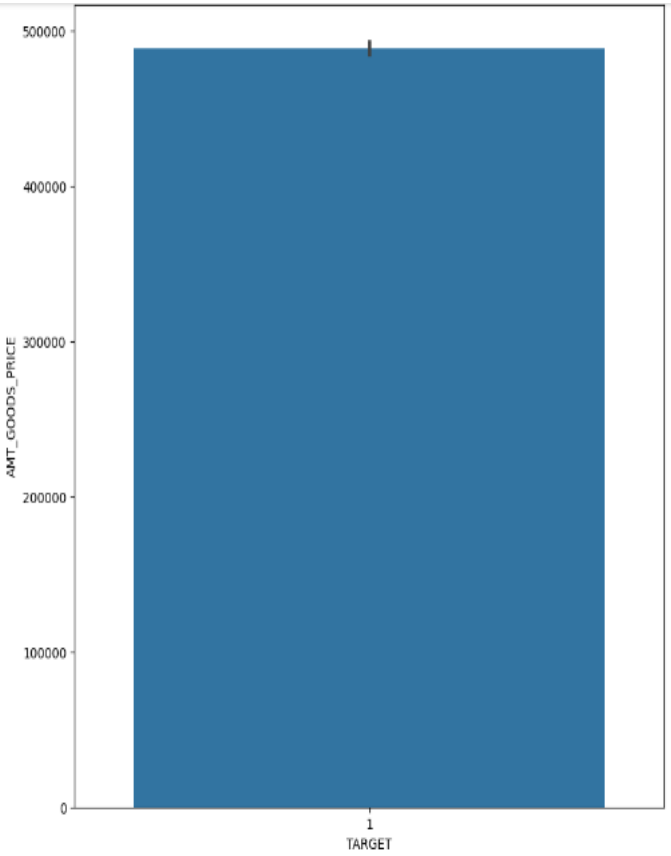
Insights from univariate analysis of Categorical variables

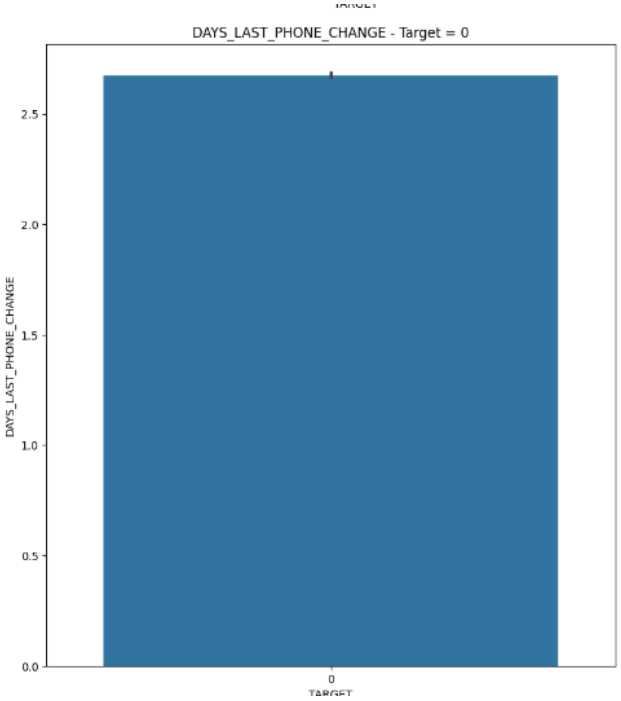
- **Code_Gender**: Defaulters (Target = 1) has a higher percentage of male customers in comparison to non-defaulters (Target = 0)
- **NAME_EDUCATION_TYPE**: Defaulters (Target = 1) has a higher percentage of customers with Secondary/Secondary Special education
- **Age_Group** : Defaulters (Target = 1) has a higher percentage of customers in the age group of 30s
- **NAME_INCOME_TYPE**: Defaulters (Target = 1) has a higher percentage of working customers whereas percentage of defaulting pensioners is lesser in comparison to non-defaulters(Target = 0)
- **OCCUPATION_TYPE**: Laborers contribute a higher percentage in defaulters(Target = 1) in comparison to non-defaulters(Target = 0)
- **CNT_CHILDREN** : : Defaulters (Target = 0) has a higher percentage 0 child in comparison to non-defaulters (Target = 0)
- **NAME_CONTRACT_TYPE**:-Case loans in repayment status and defaulting status in quite the same
- **FLAG_OWN_CAR**:- Clients having no car have more defaulter rate.

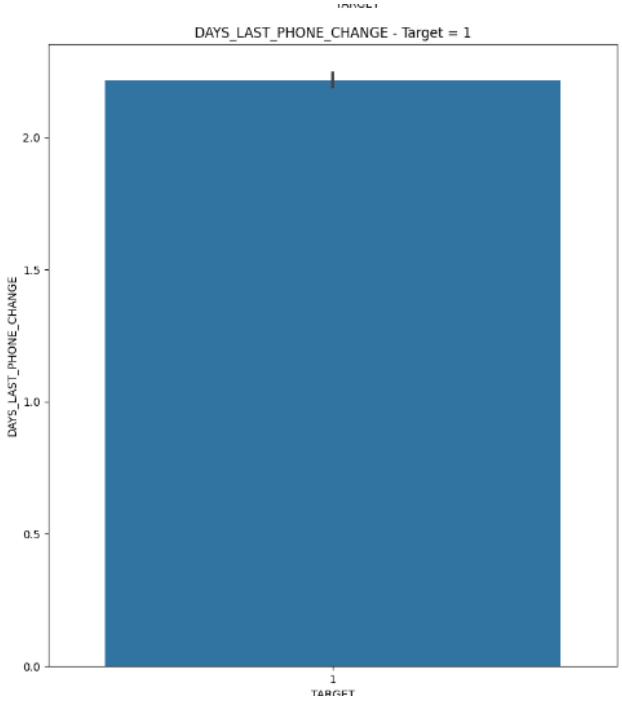
The objective of this analysis is to understand how numerical variables vary between target 0 and 1

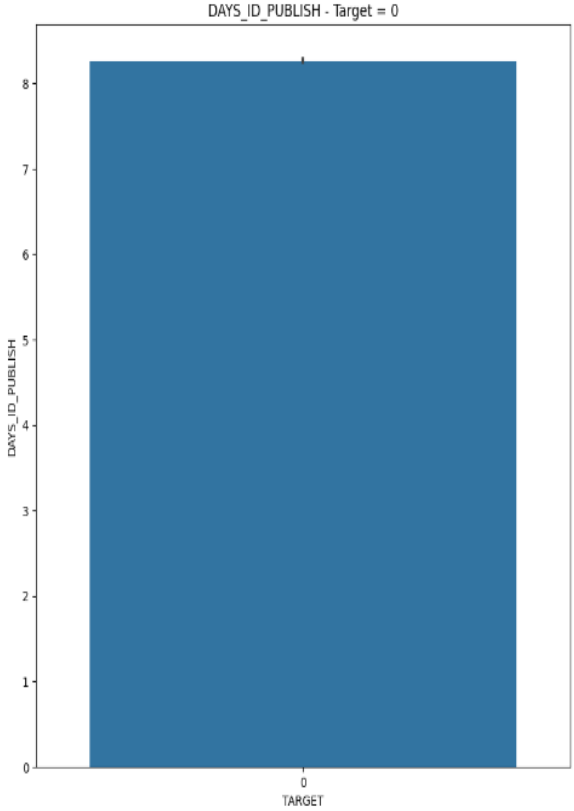
```
list of all continuous numerical column
numerical_columns= ['AMT_GOODS_PRICE',
'DAYS_LAST_PHONE_CHANGE',
'DAYS_ID_PUBLISH',
'AMT_INCOME_TOTAL',
'DAYS_EMPLOYED',
'DAYS_REGISTRATION',
'DAYS_BIRTH',
'AMT_CREDIT',
'AMT_ANNUITY'
]
```

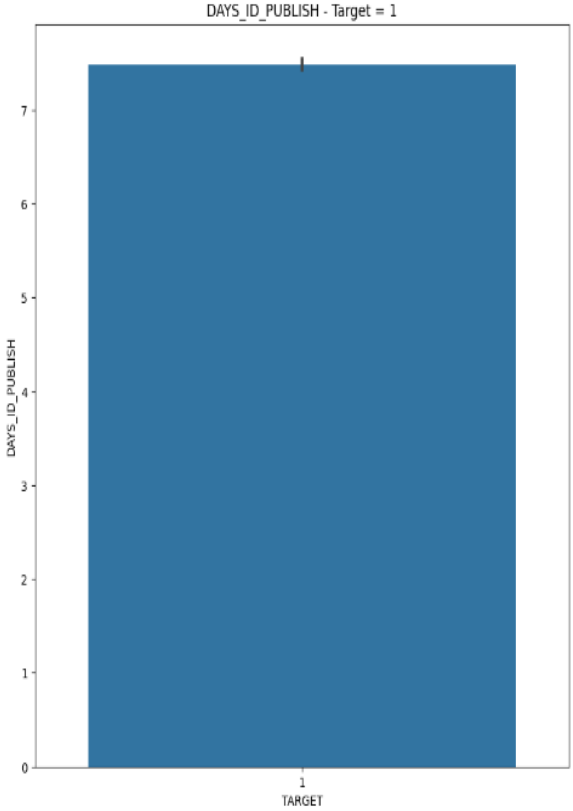


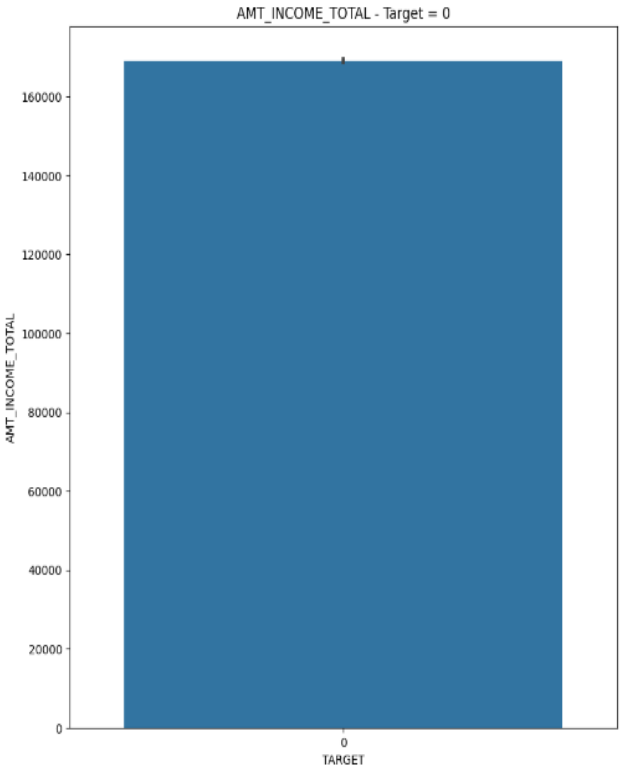


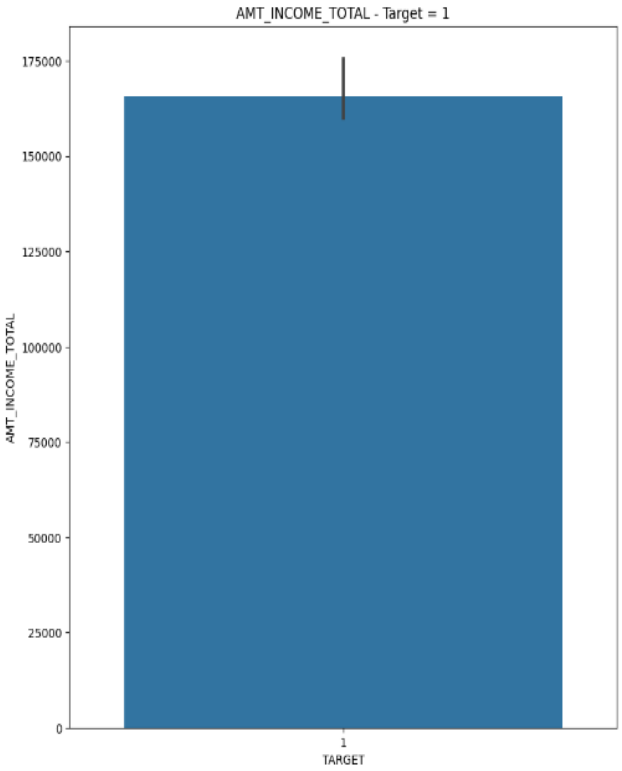

DAYS_LAST_PHONE_CHANGE - Target = 0
DAYS_LAST_PHONE_CHANGE
0.0
0.5
1.0
1.5
2.0
2.5
0
TARGET

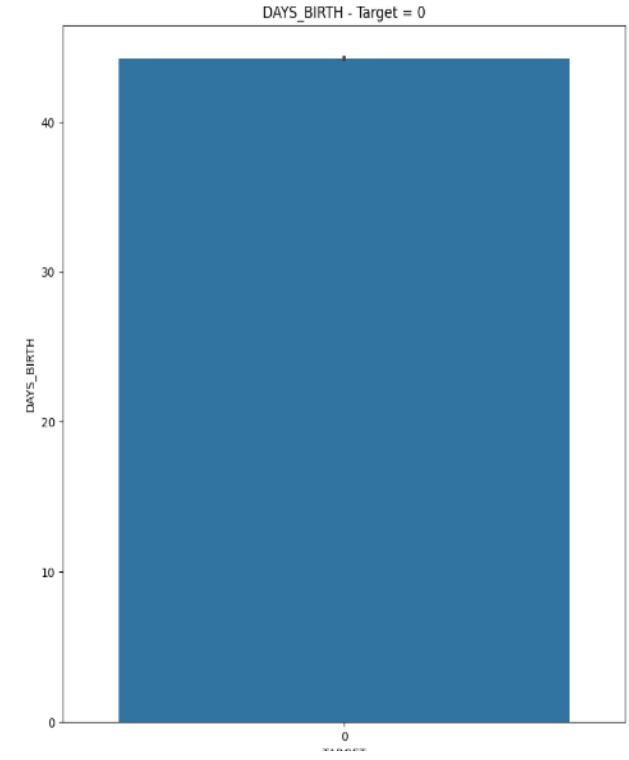

DAYS_LAST_PHONE_CHANGE - Target = 1
DAYS_LAST_PHONE_CHANGE
0.0
0.5
1.0
1.5
2.0
1
TARGET

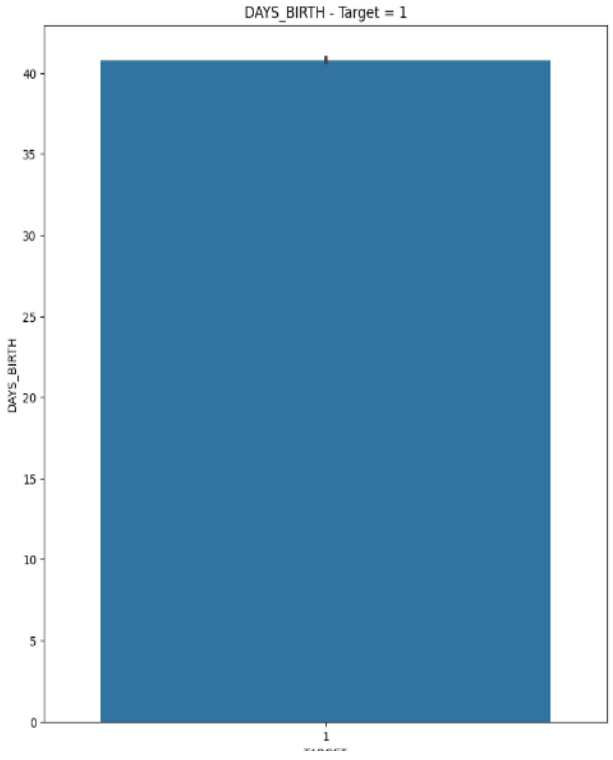

DAYS_ID_PUBLISH - Target = 0
DAYS_ID_PUBLISH
0
1
2
3
4
5
6
7
8
0
TARGET

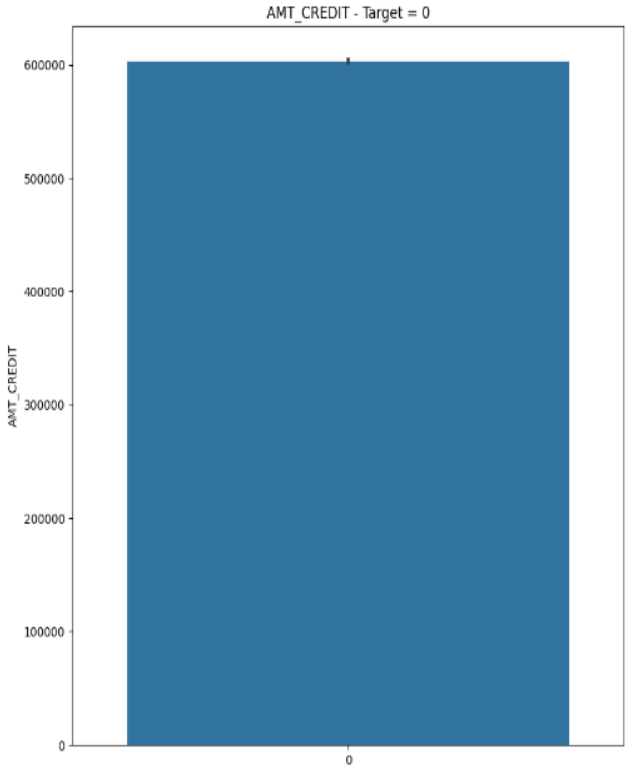

DAYS_ID_PUBLISH - Target = 1
DAYS_ID_PUBLISH
0
1
2
3
4
5
6
7
1
TARGET

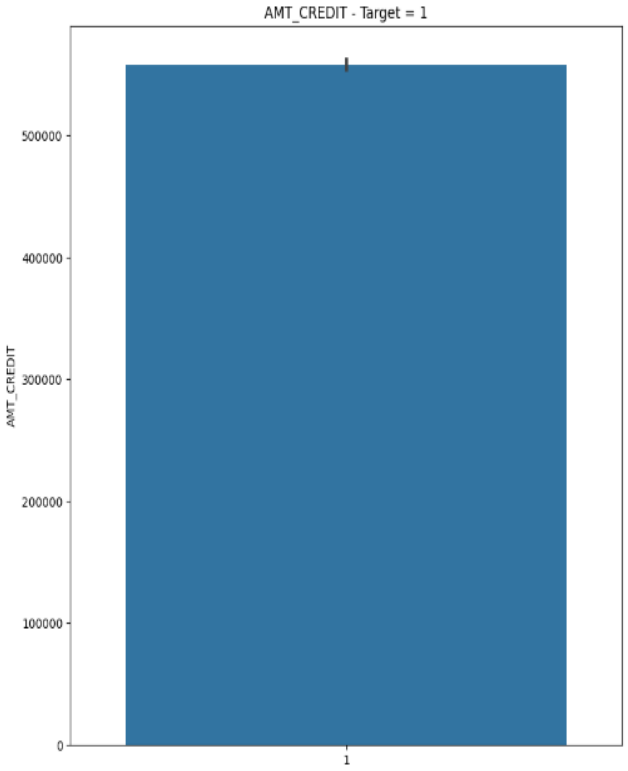

AMT_INCOME_TOTAL - Target = 0
AMT_INCOME_TOTAL
TARGET
0
0
20000
40000
60000
80000
100000
120000
140000
160000

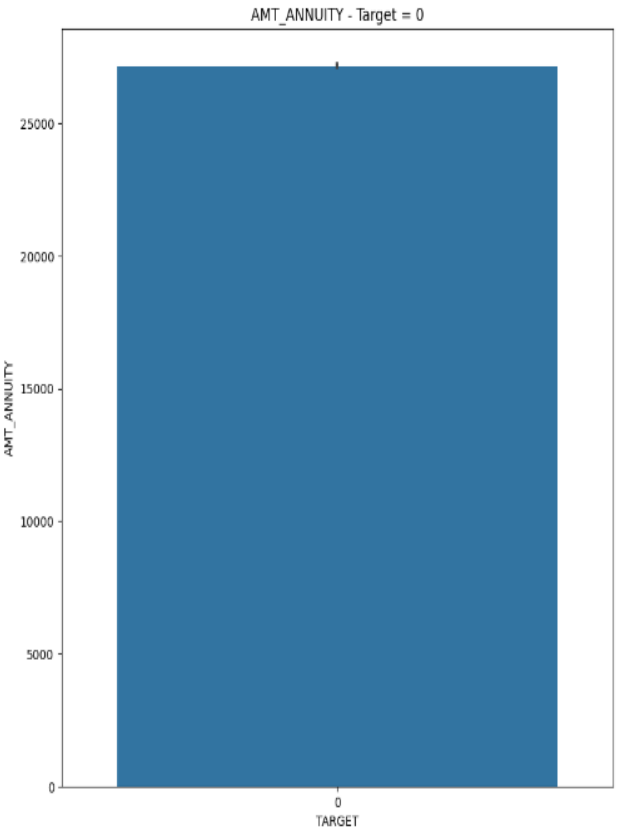

AMT_INCOME_TOTAL - Target = 1
AMT_INCOME_TOTAL
TARGET
1
0
25000
50000
75000
100000
125000
150000
175000

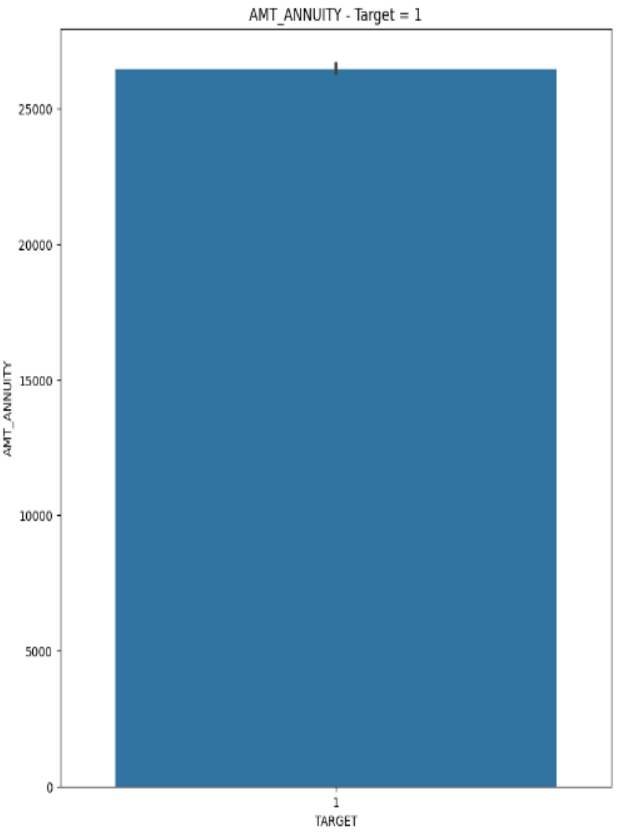

DAYS_BIRTH - Target = 0
DAYS_BIRTH
0
0
10
20
30
40


DAYS_BIRTH - Target = 1
DAYS_BIRTH
1
0
5
10
15
20
25
30
35
40


AMT_CREDIT - Target = 0
AMT_CREDIT
600000
500000
400000
300000
200000
100000
0
0


AMT_CREDIT - Target = 1
AMT_CREDIT
500000
400000
300000
200000
100000
0
1

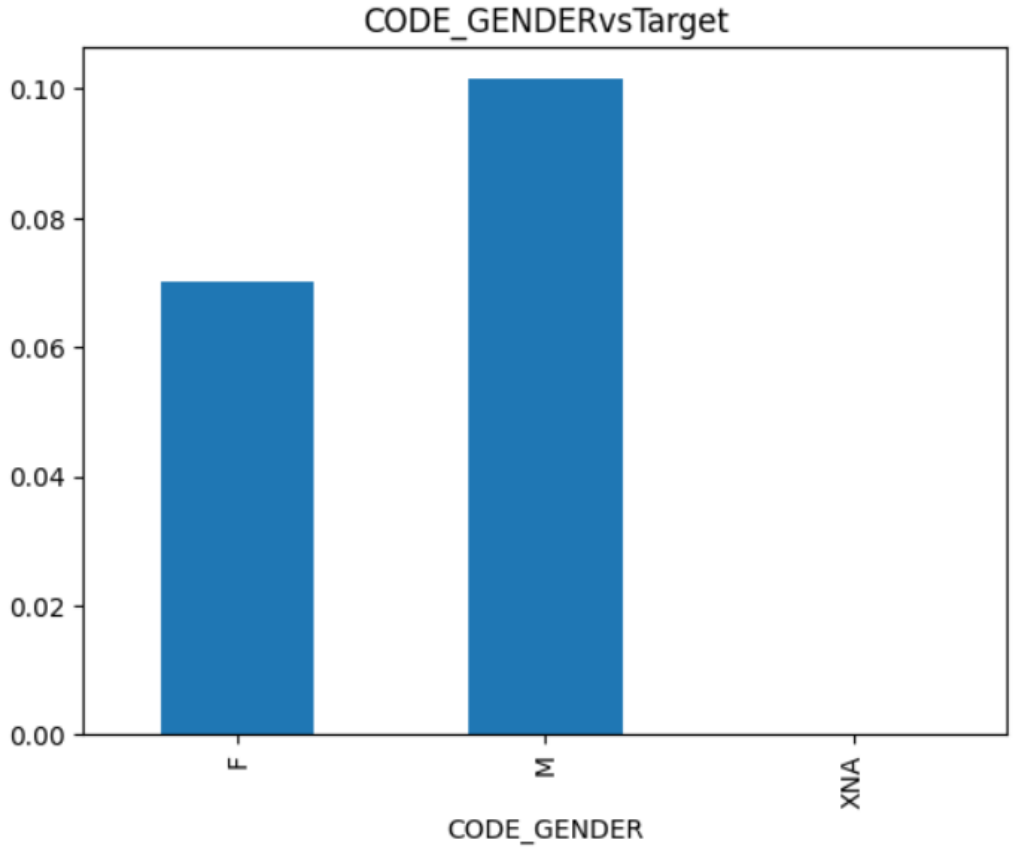

AMT_ANNUITY - Target = 0
AMT_ANNUITY
25000
20000
15000
10000
5000
0
0
TARGET

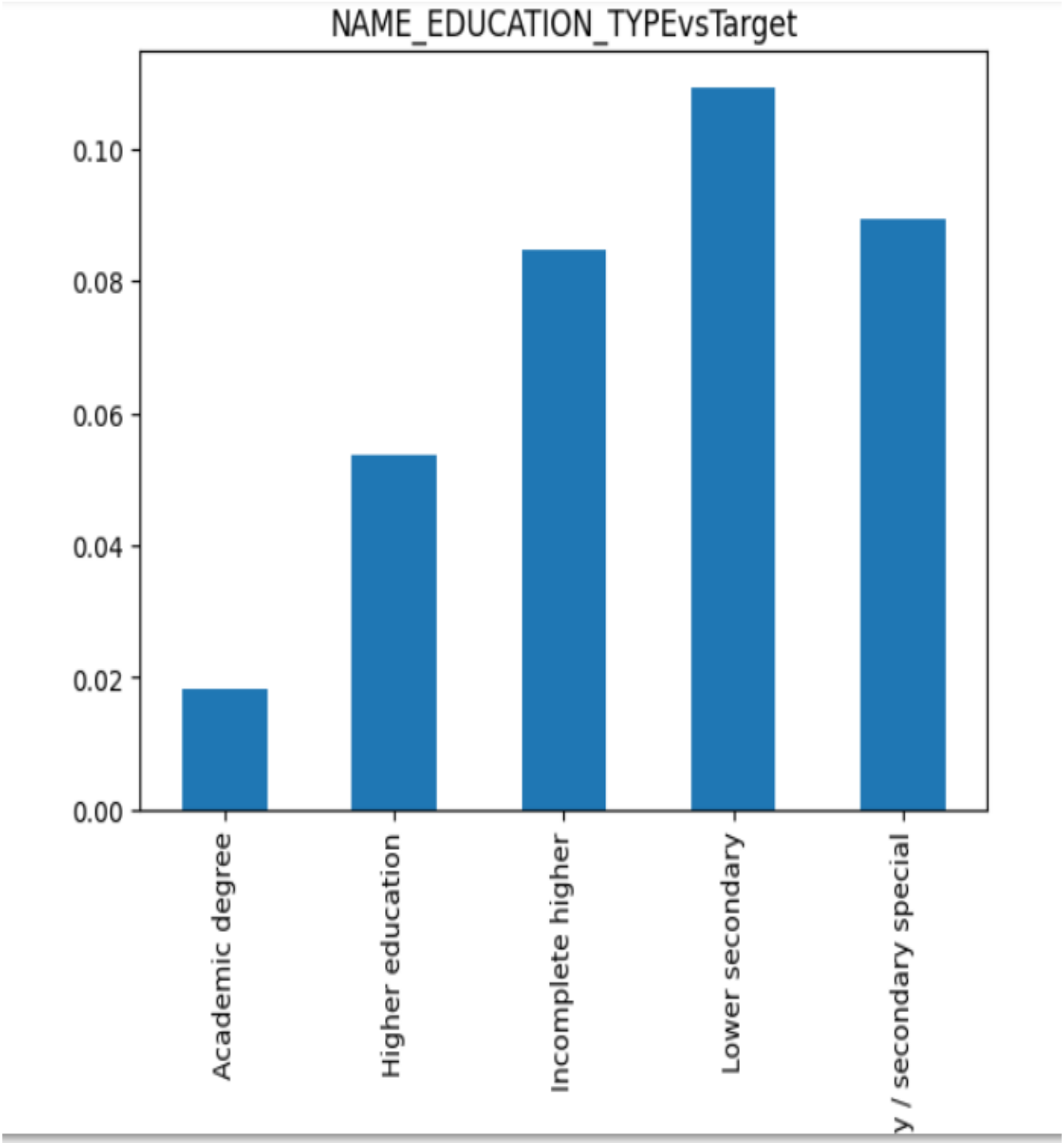

AMT_ANNUITY - Target = 1
AMT_ANNUITY
25000
20000
15000
10000
5000
0
1
TARGET

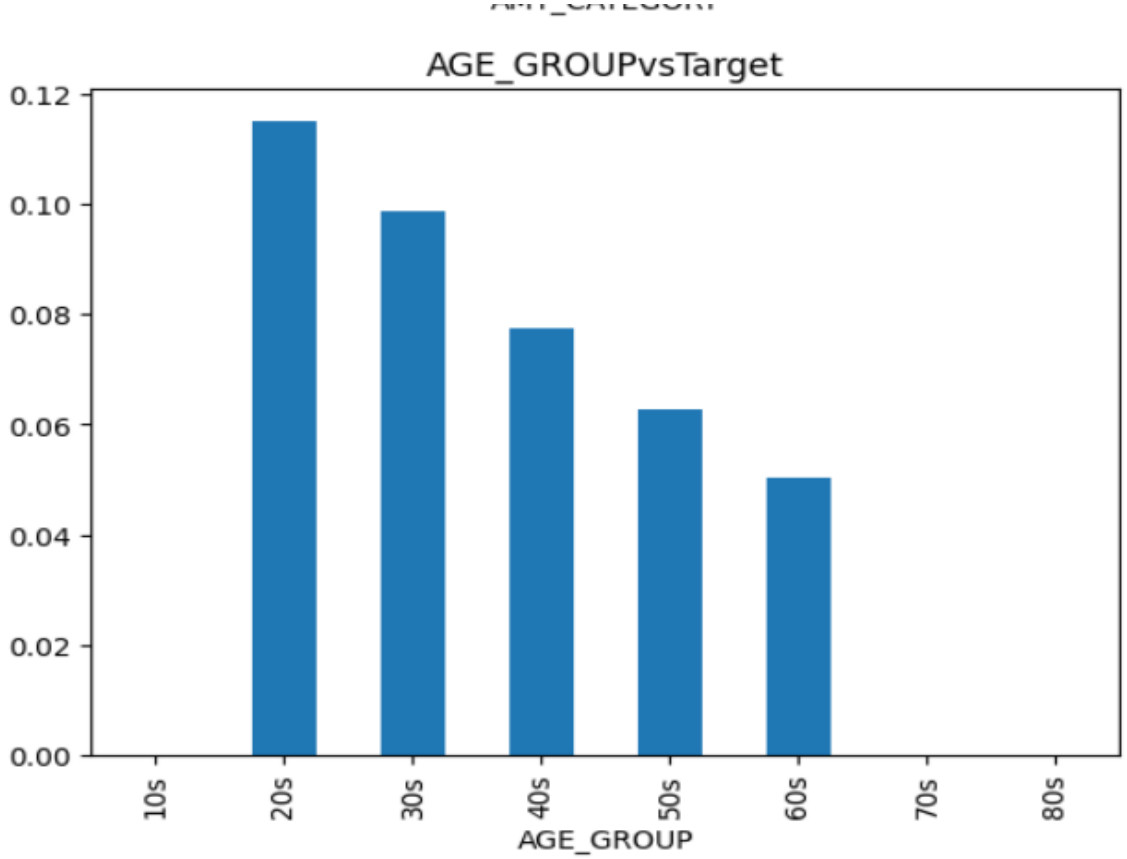
Insights from univariate analysis of numerical variables

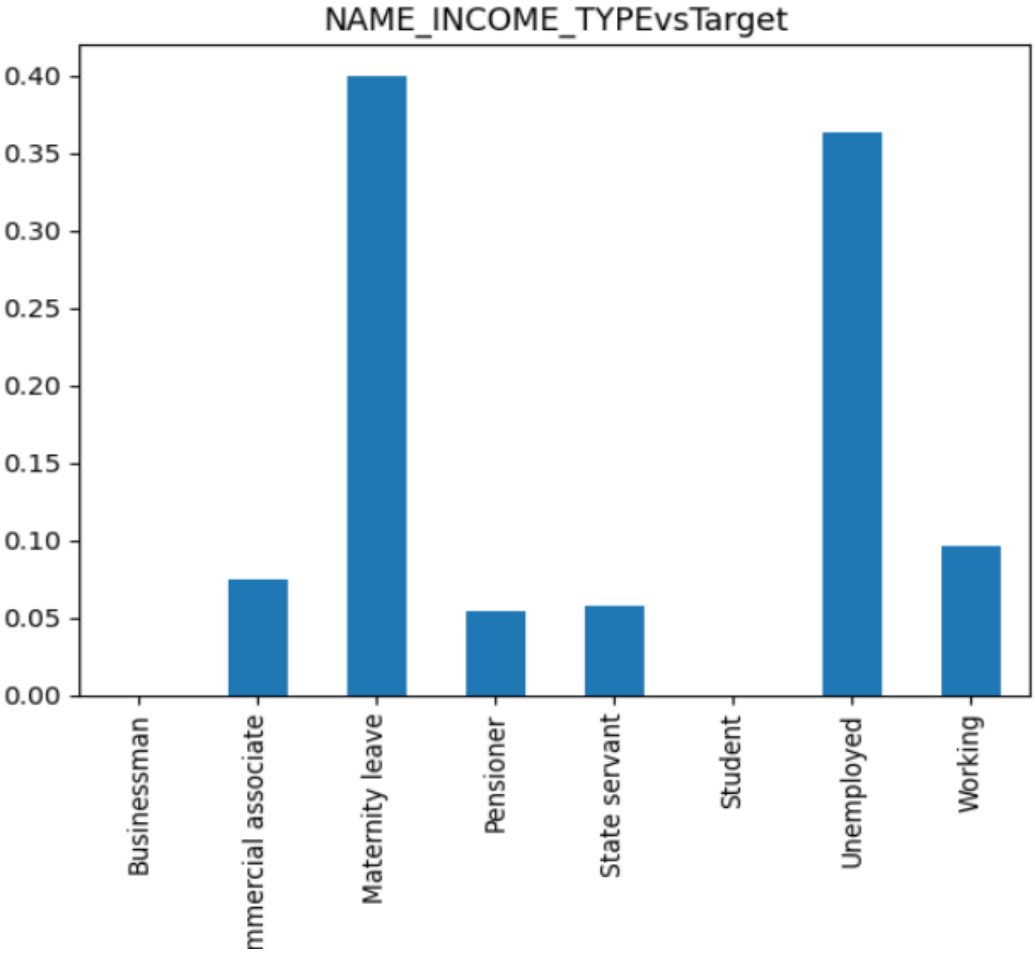
- **DAYS_LAST_PHONE_CHANGE**. It implies non defaulter more often change phone number before application
- **DAYS_ID_PUBLISH**: non Defaulters seem to change IDs more frequently than defaulters
- **DAYS_BIRTH**: defaulter population is younger than non-defaulter.
- **AMT_ANNUITY**:-In both the cases, it is almost same.
- **AMT_INCOME_TOTAL** =The people having income greater than or equal to 1.6 Lakh they are in more number for returning loans as compared to defaulter status.

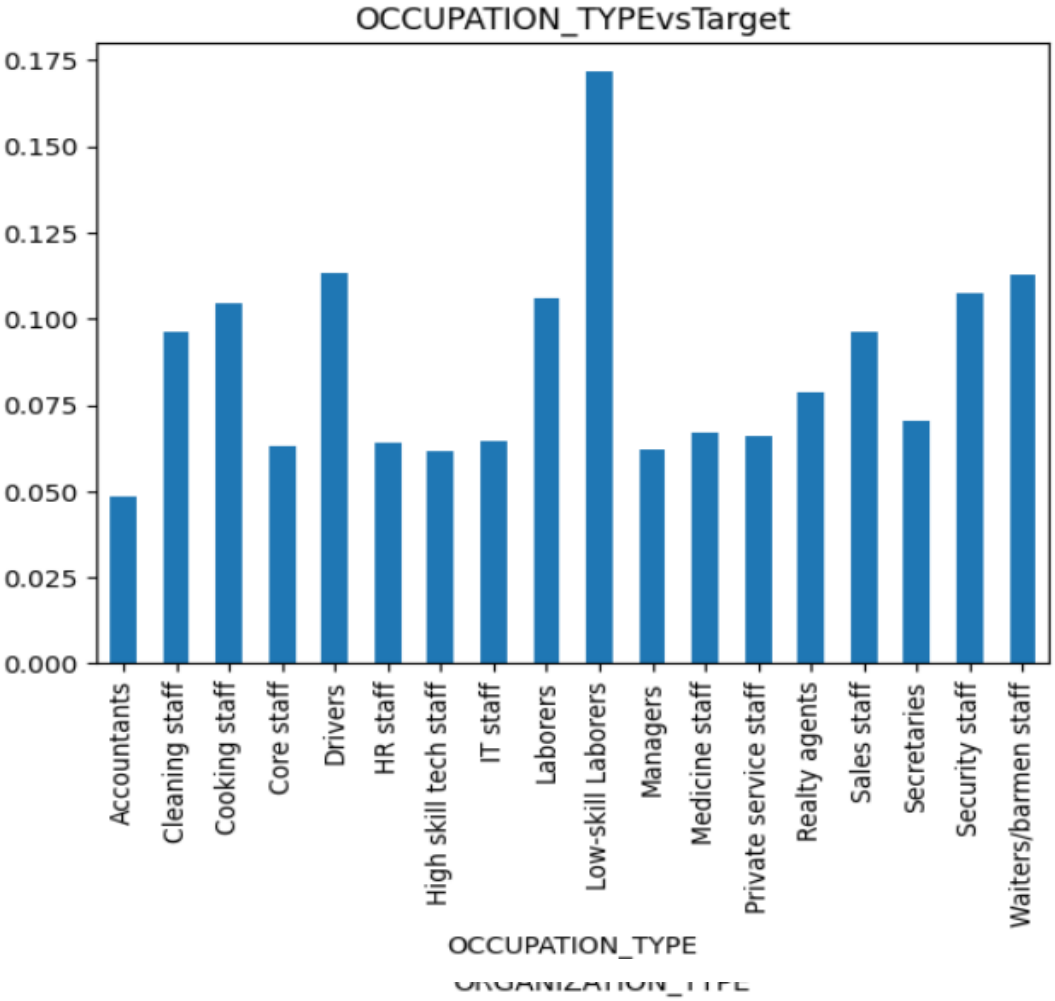
Bivariate Analysis for Categorical Variable

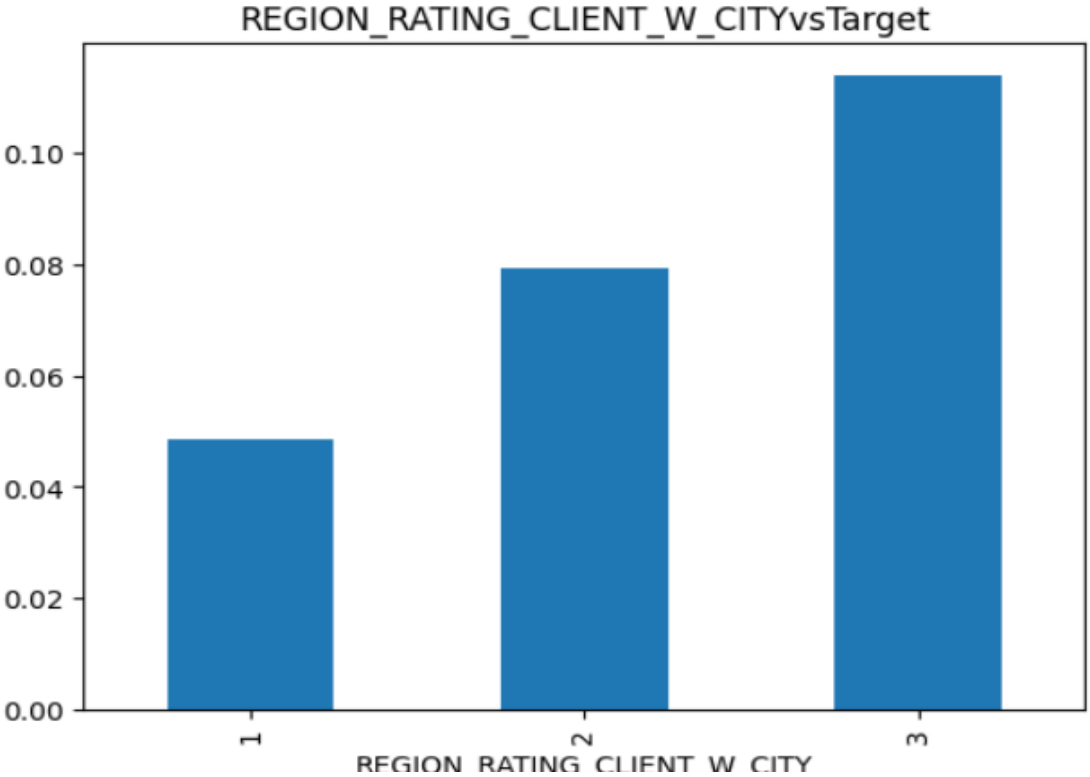


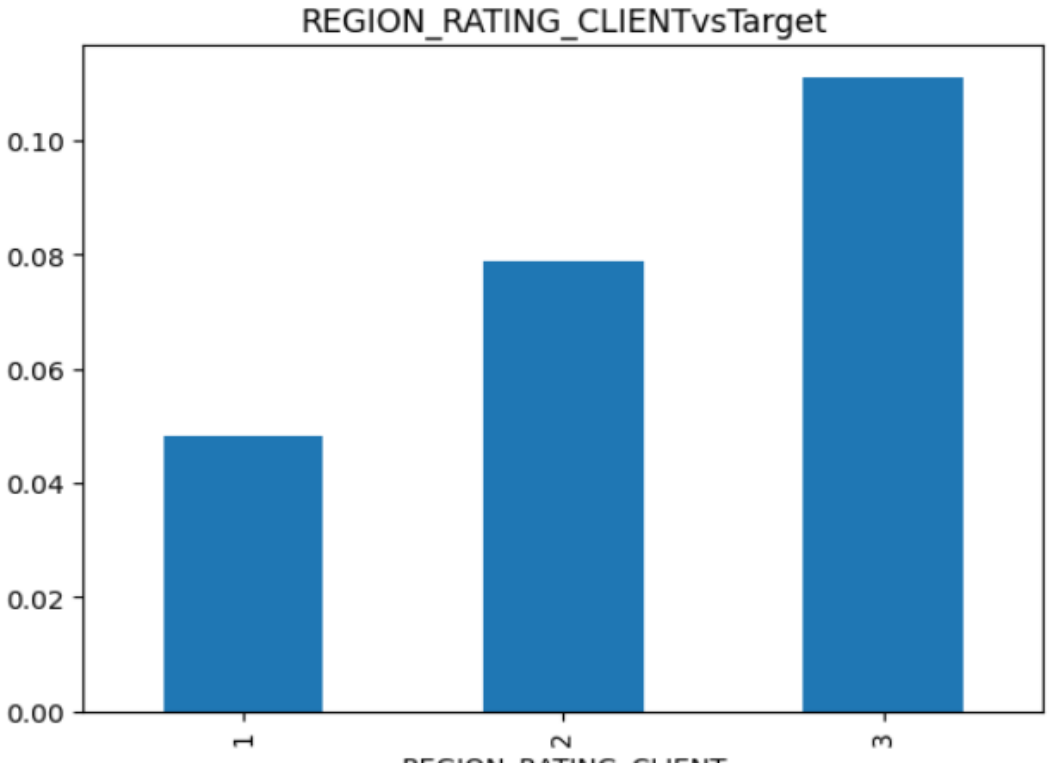

NAME_EDUCATION_TYPEvsTarget
0.10
0.08
0.06
0.04
0.02
0.00
Academic degree
Higher education
Incomplete higher
Lower secondary
y / secondary special

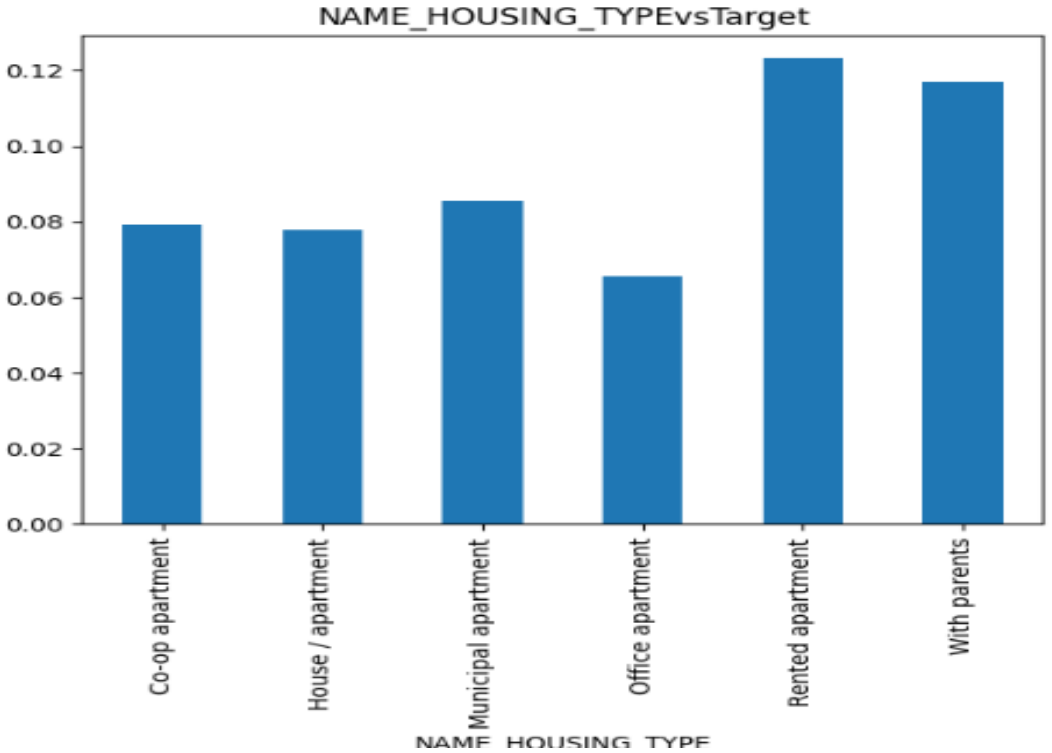

AGE_GROUPvsTarget
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10s
20s
30s
40s
50s
60s
70s
80s
AGE_GROUP


NAME_INCOME_TYPEvsTarget
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Businessman
mmercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working


OCCUPATION_TYPEvsTarget
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
HR staff
High skill tech staff
IT staff
Laborers
Low-skill Laborers
Managers
Medicine staff
Private service staff
Realty agents
Sales staff
Secretaries
Security staff
Waiters/barmen staff
OCCUPATION_TYPE


REGION_RATING_CLIENT_W_CITYvsTarget
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3

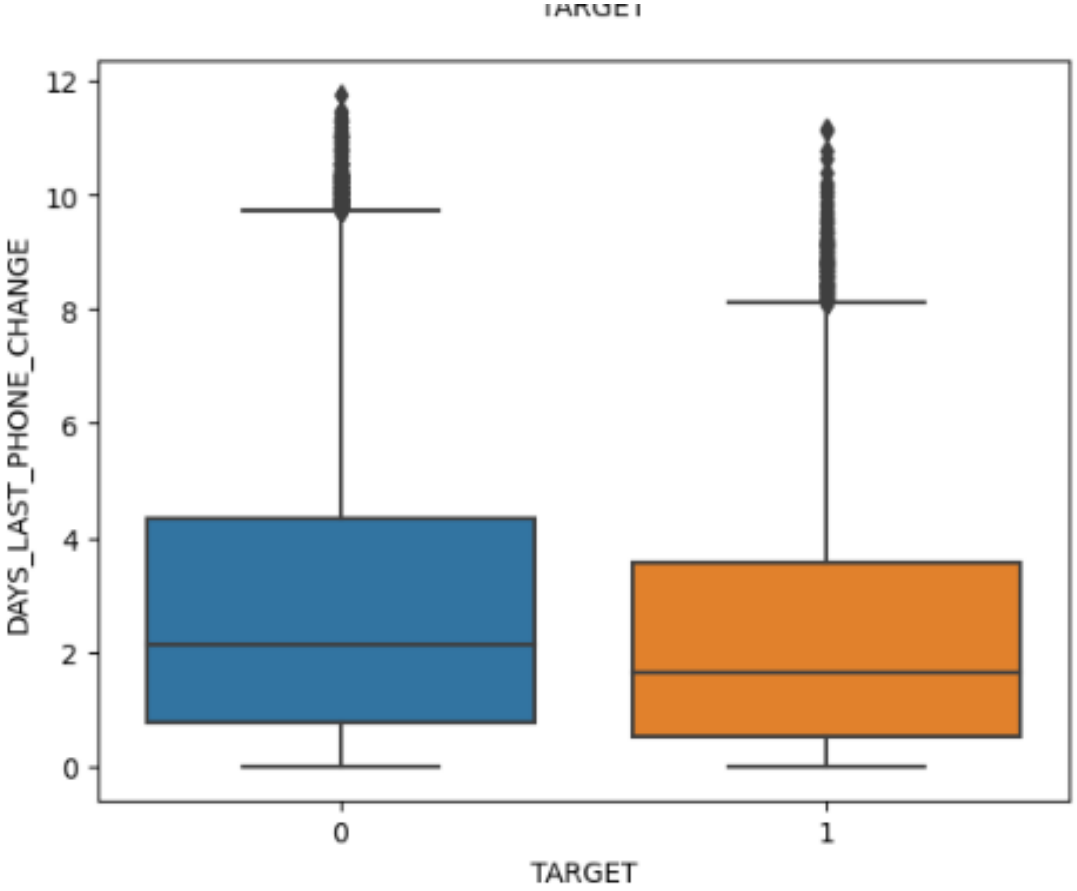

REGION_RATING_CLIENTvsTarget
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
NAME_FAMILY_STATUS

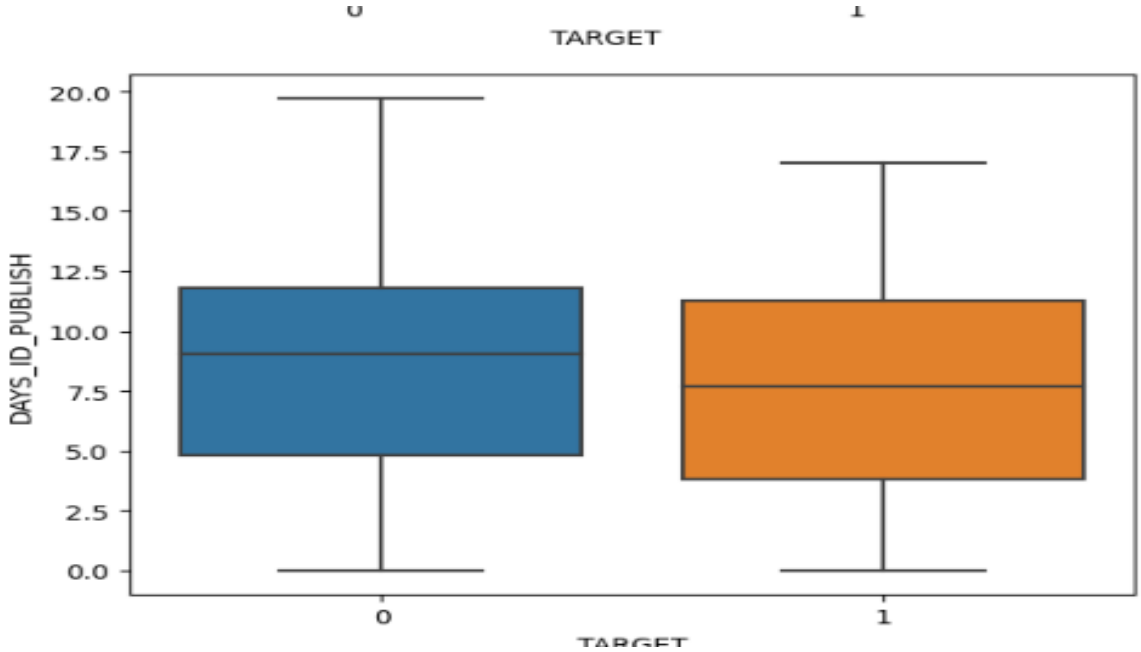

NAME_HOUSING_TYPEvsTarget
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents

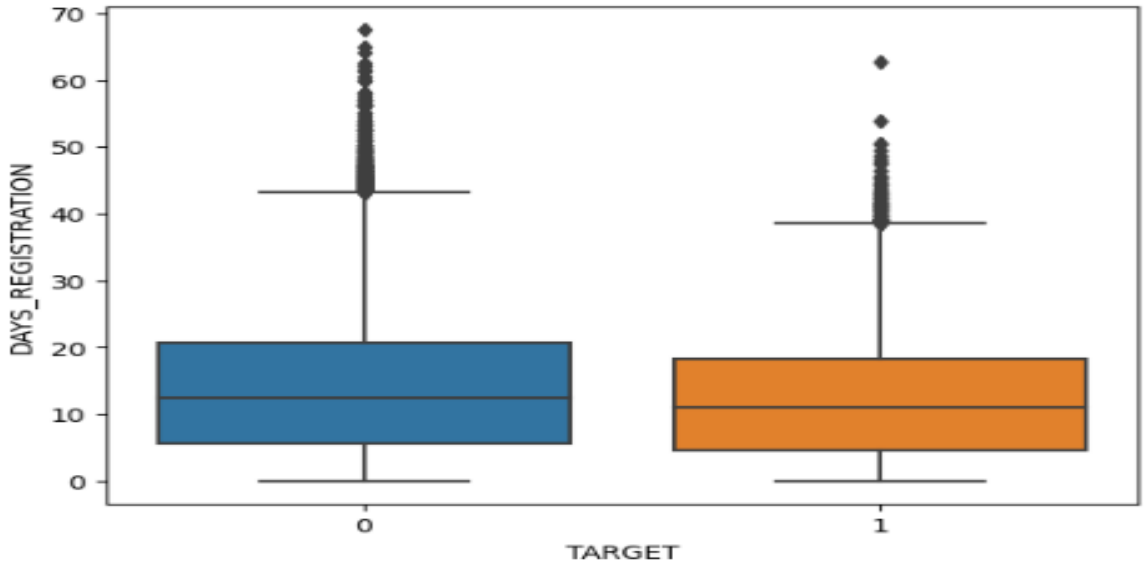
insights from bivariate analysis of categorical variables

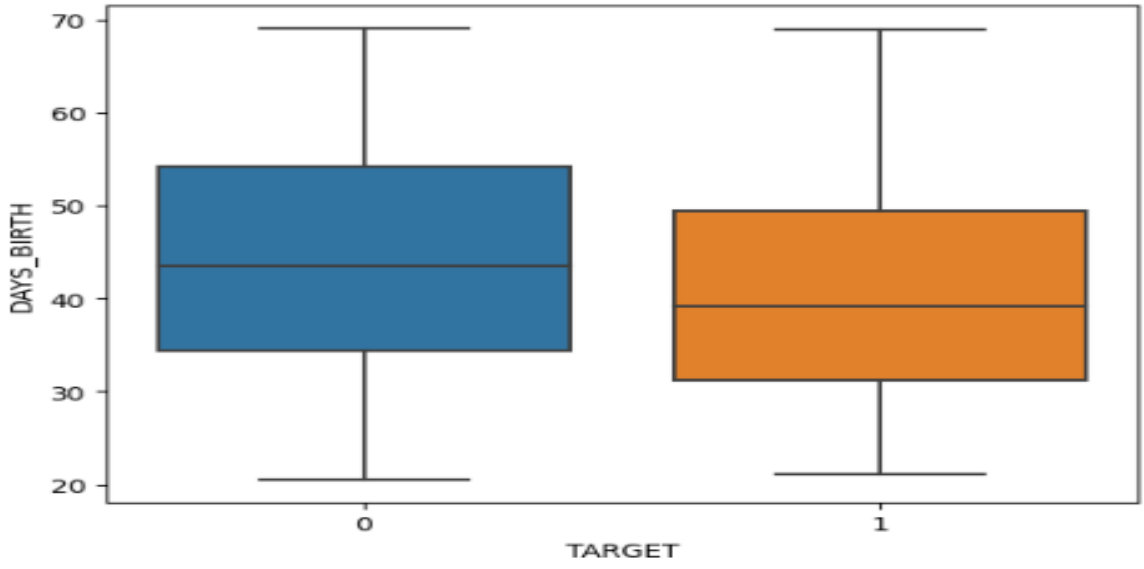
- **CODE_GENDER**: Male customers have a higher probability of defaulting
- **NAME_EDUCATION_TYPE**:Customers with lower secondary education have a higher risk of default
- **AGE_GROUP**- Customers in 20s and 30s have higher chances of defaulting
- **NAME_HOUSING_TYPE**: Customers living in rented apartments and living with parents seem to default more
- **NAME_INCOME_TYPE**:Unemployed and Customers on maternity leave have higher
- **OCCUPATION_TYPE**:Low-skill laborers default more
- **REGION_RATING_CLIENT&REGION_RATING_CLIENT_W_CITY**: Customers with rating 3 have higher risk of defaulting
- People who get income through Maternity Leave tend to be more Defaulter when they have more Family Members

Bivariate Analsyis for Numerical columns




TARGET
DAYS_ID_PUBLISH
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1


DAYS_REGISTRATION
70
60
50
40
30
20
10
0
0
1
TARGET


DAYS_BIRTH
70
60
50
40
30
20
0
1
TARGET

In this section, we will only highlight key outcomes from Bivariate analysis

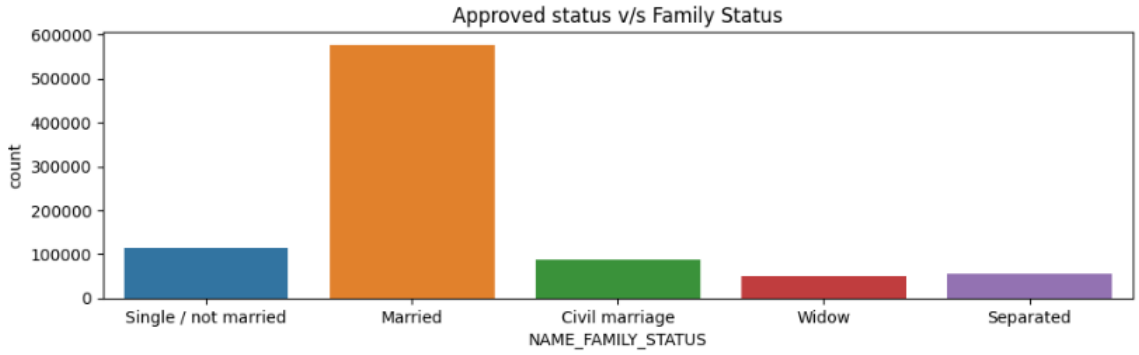
- DAYS_LAST_PHONE_CHANGE:Defaulter customers change phone closer to the submission of application
- DAYS ID_PUBLISH:Defaulter customers changes id closer to submission of application
- DAYS_REGISTRATION: Defaulter customers changes registration on a date closer to submission of application
- DAYS_BIRTH: Defaulter customers are relatively younger than non-defaulters

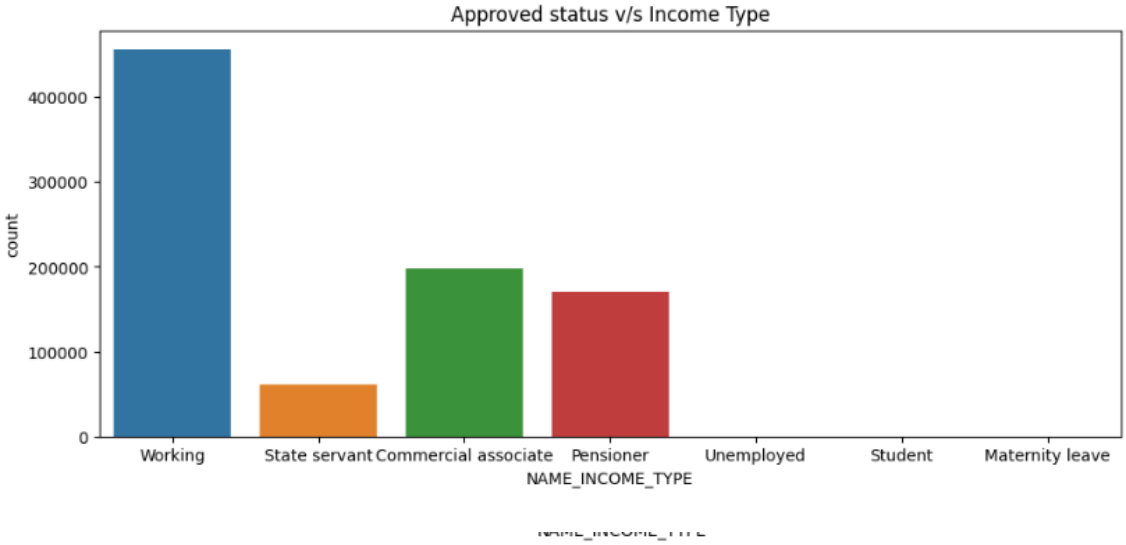
There are four types of decisions that could be taken by the client/company:

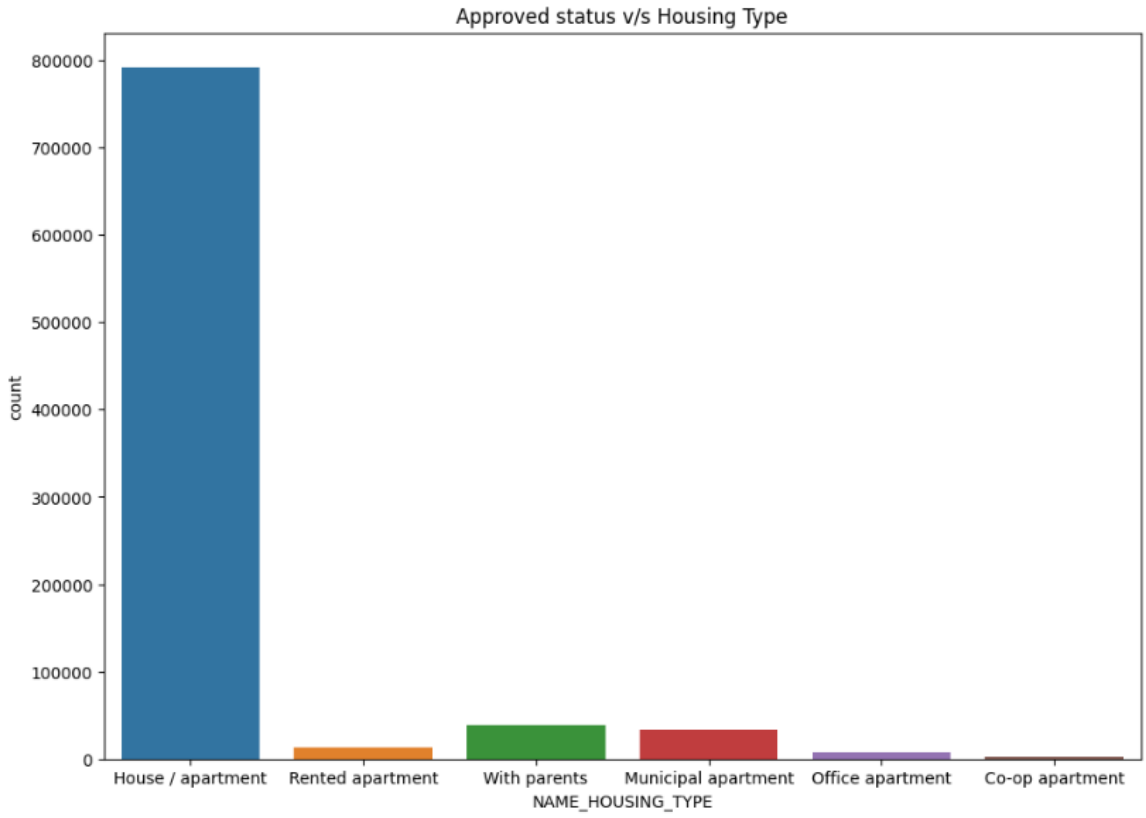
1. **Approved:** The company has approved loan application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused Offer:** Loan has been cancelled by the client but on different stages
5. of the process.

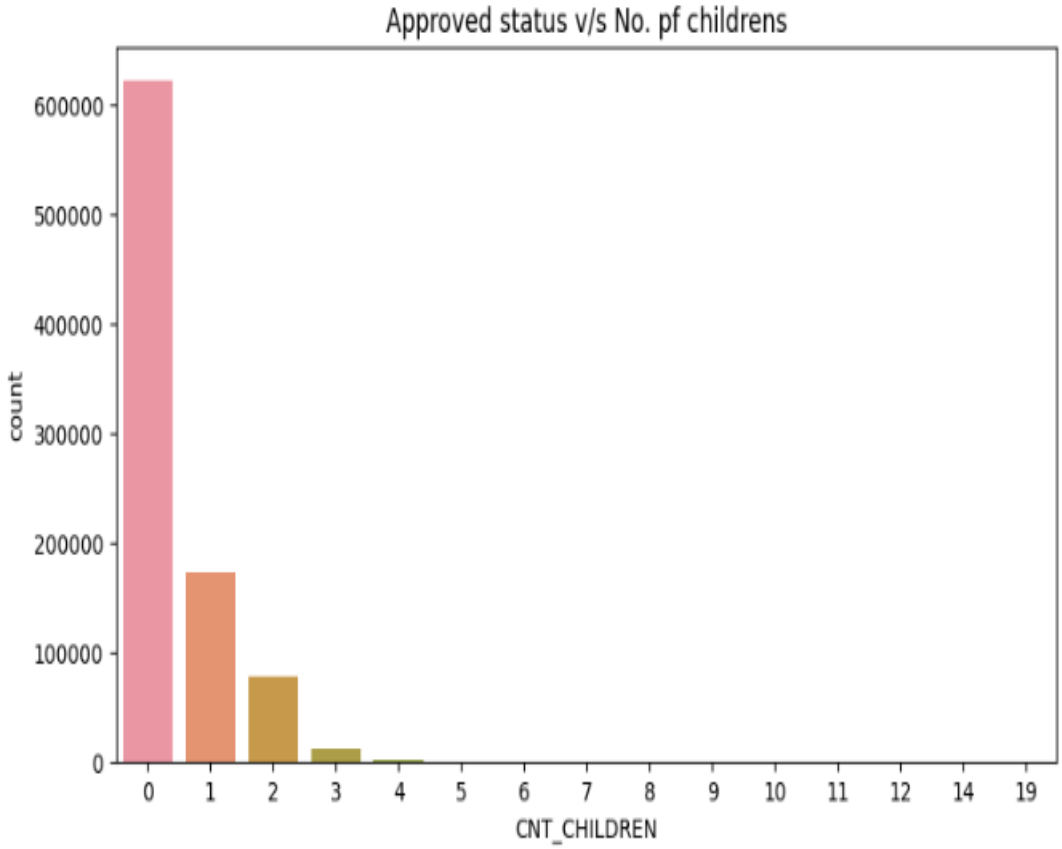
I COMBINED APPLICATION DATA AND PREVIOUS DATA (MERGED THEM)TO GAIN INSIGHTS/ANALYSIS ON THE STATUSES.

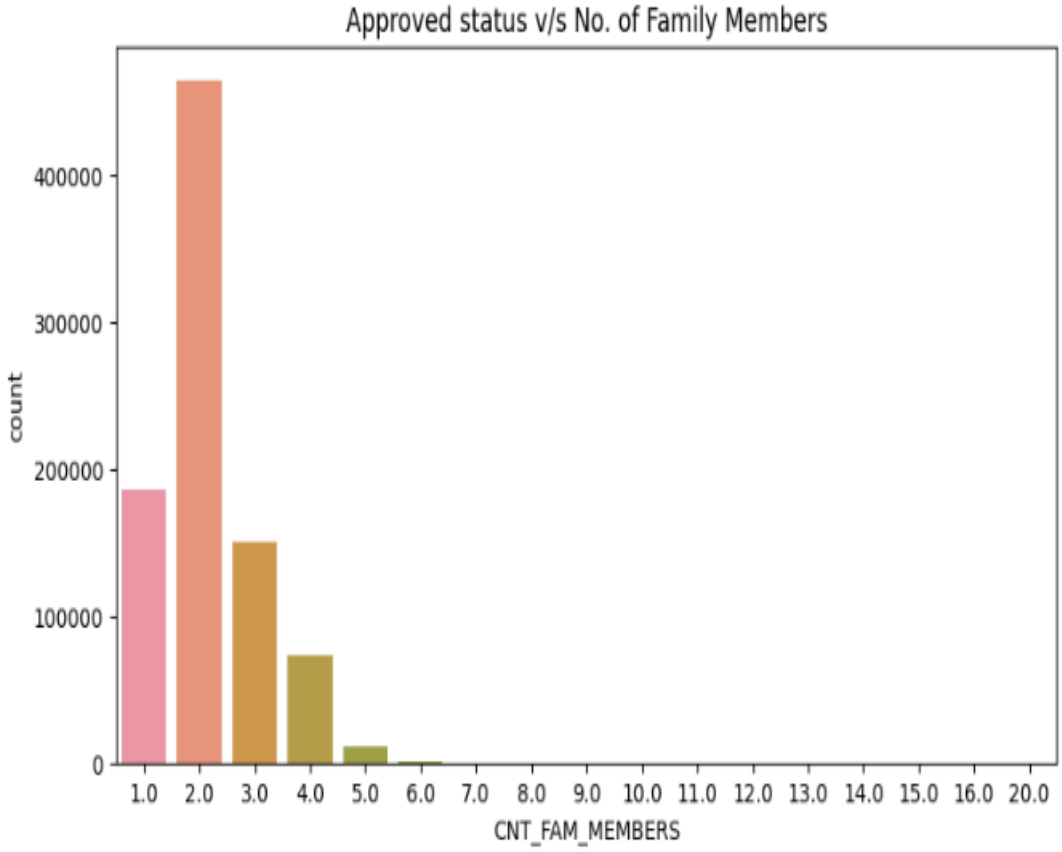
STATUS_----------APPROVED

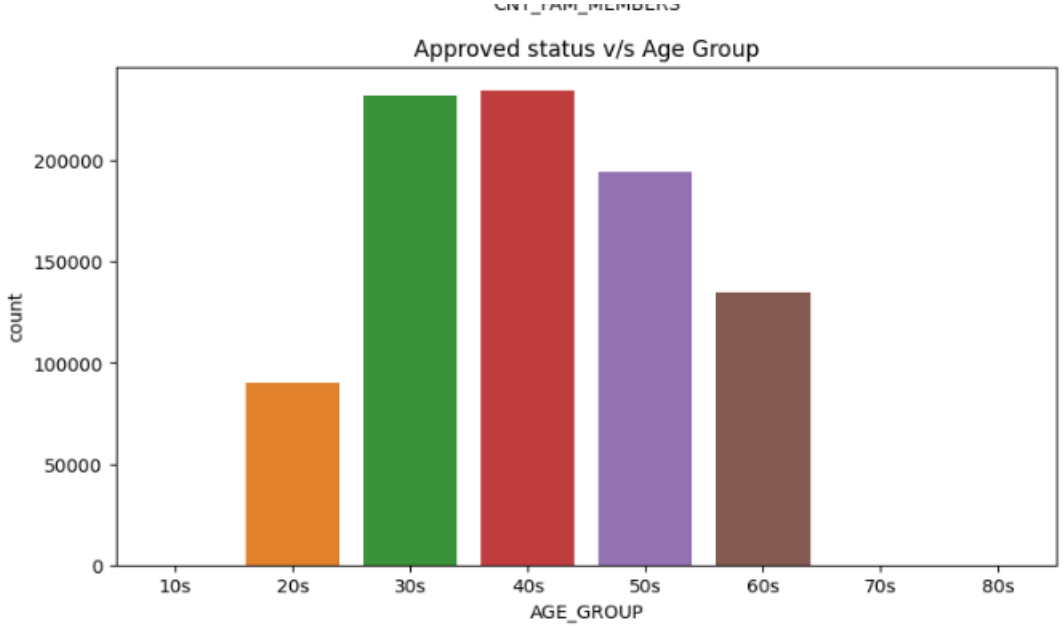



Approved status v/s Income Type
count
400000
300000
200000
100000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Maternity leave
NAME_INCOME_TYPE

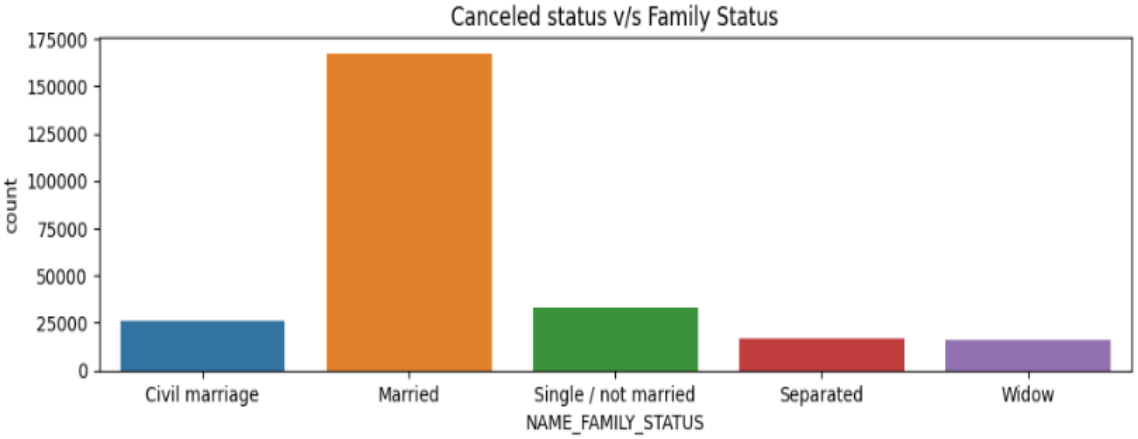

Approved status v/s Housing Type
count
800000
700000
600000
500000
400000
300000
200000
100000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

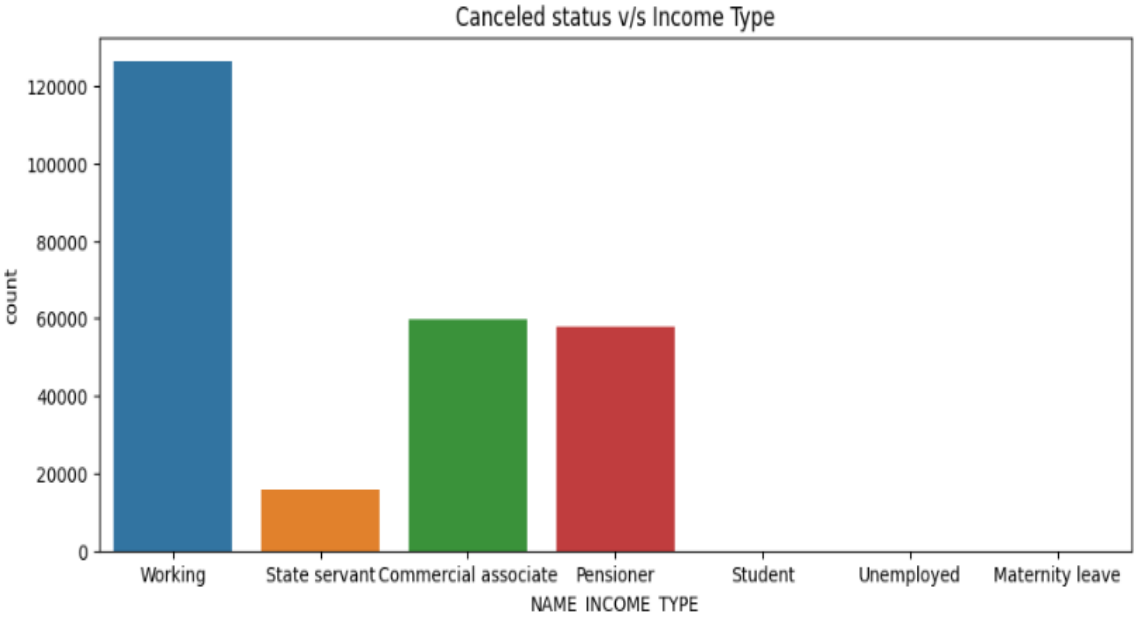

Approved status v/s No. pf childrens
count
600000
500000
400000
300000
200000
100000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 14 19
CNT_CHILDREN

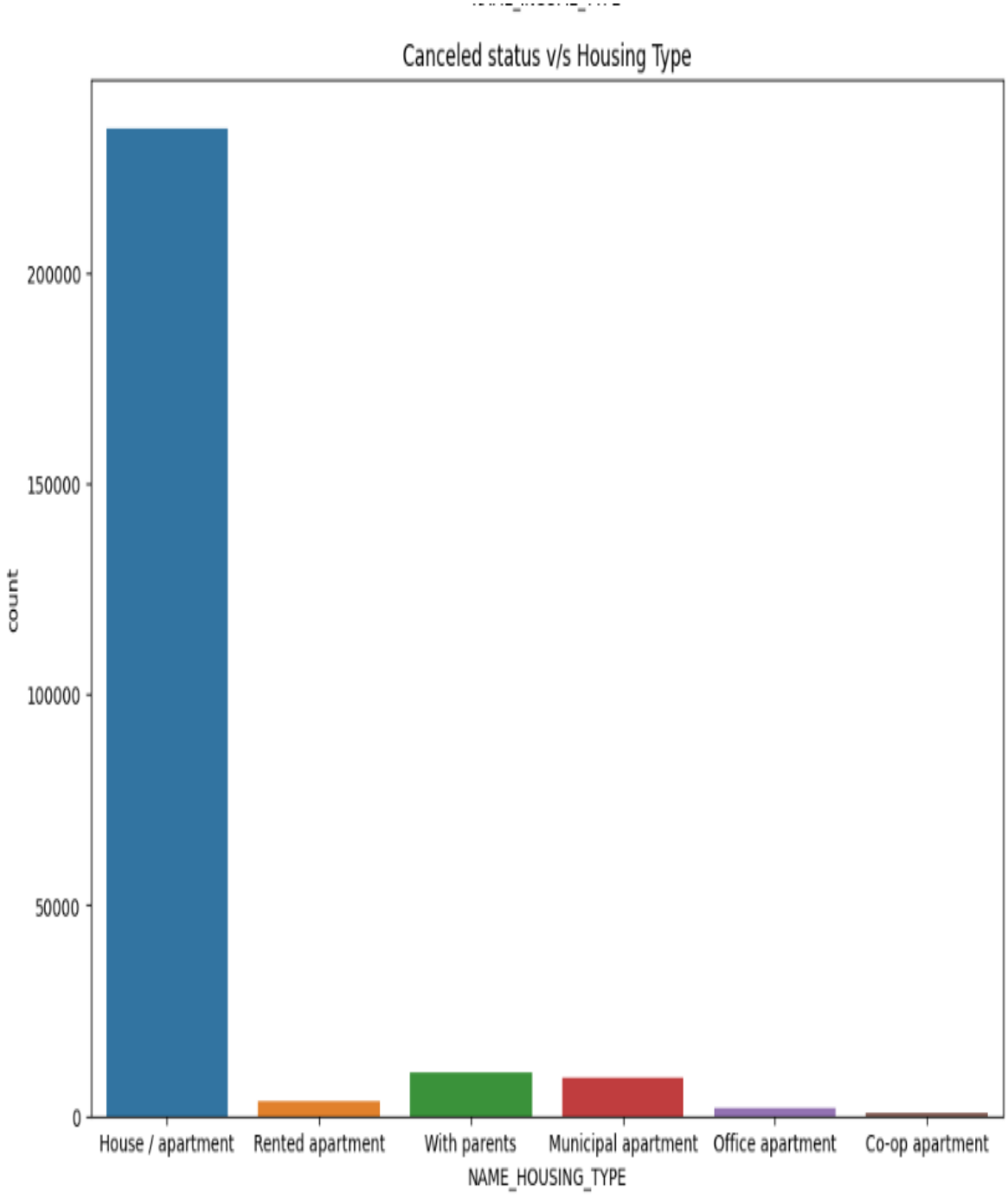

Approved status v/s No. of Family Members
count
400000
300000
200000
100000
0
1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 20.0
CNT_FAM_MEMBERS

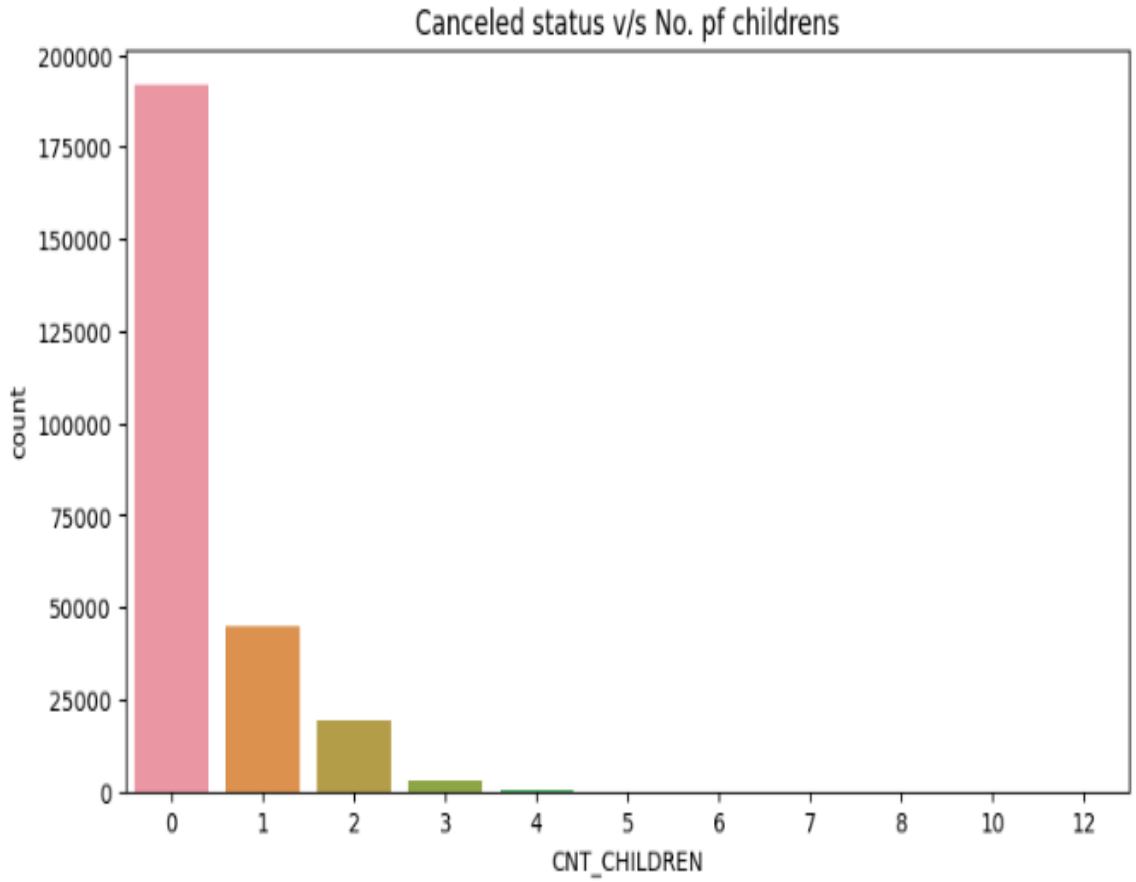


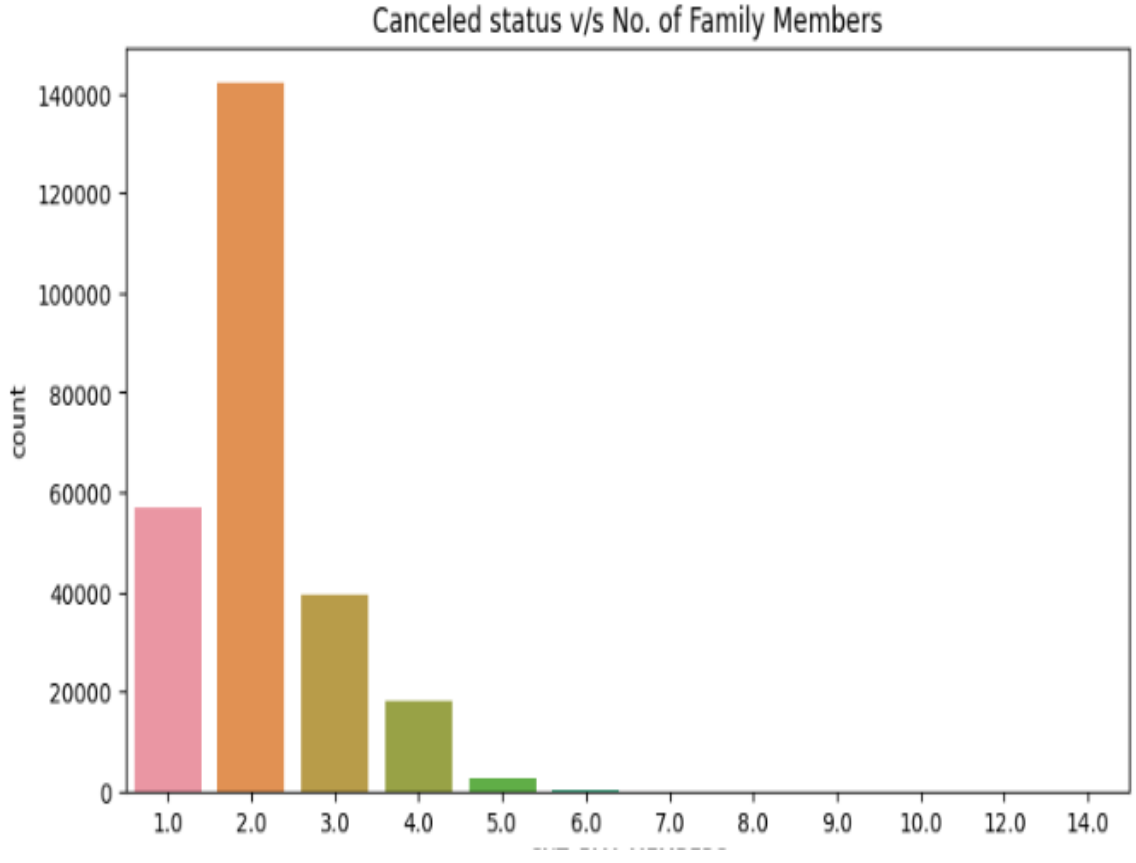
STATUS---------CANCELLED

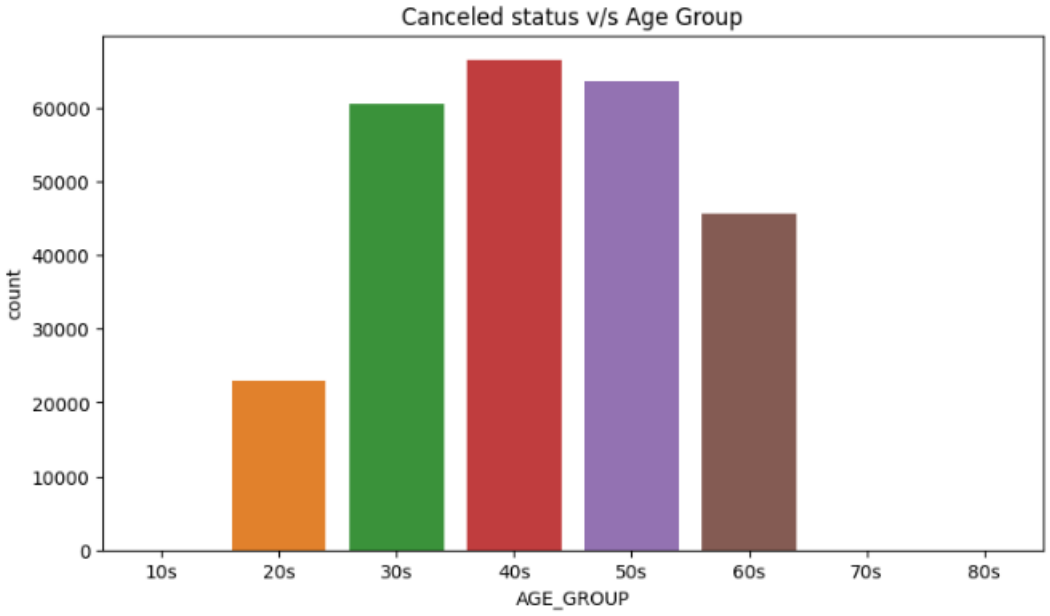




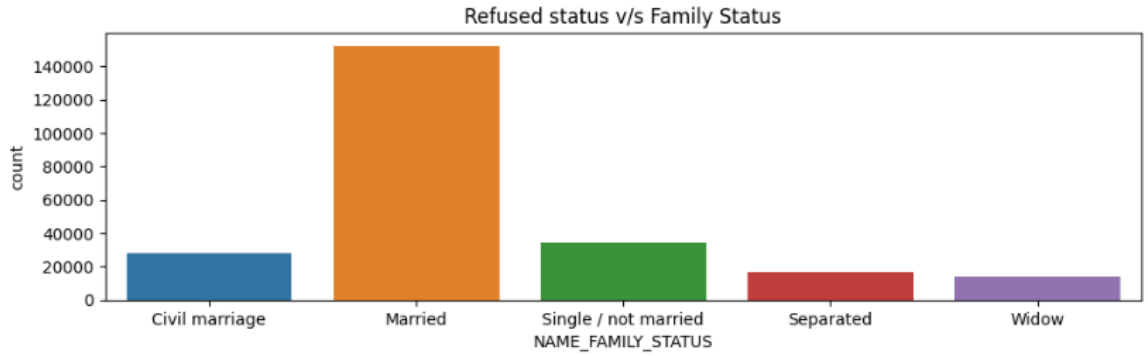

Canceled status v/s Housing Type
200000
150000
100000
50000
0
count
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

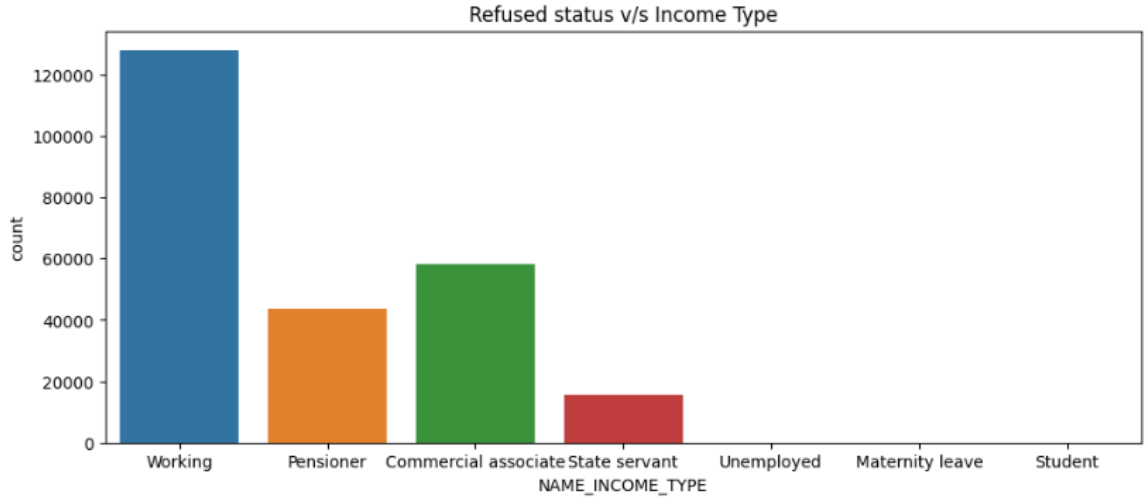

Canceled status v/s No. pf childrens
count
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
4
5
6
7
8
10
12
CNT_CHILDREN

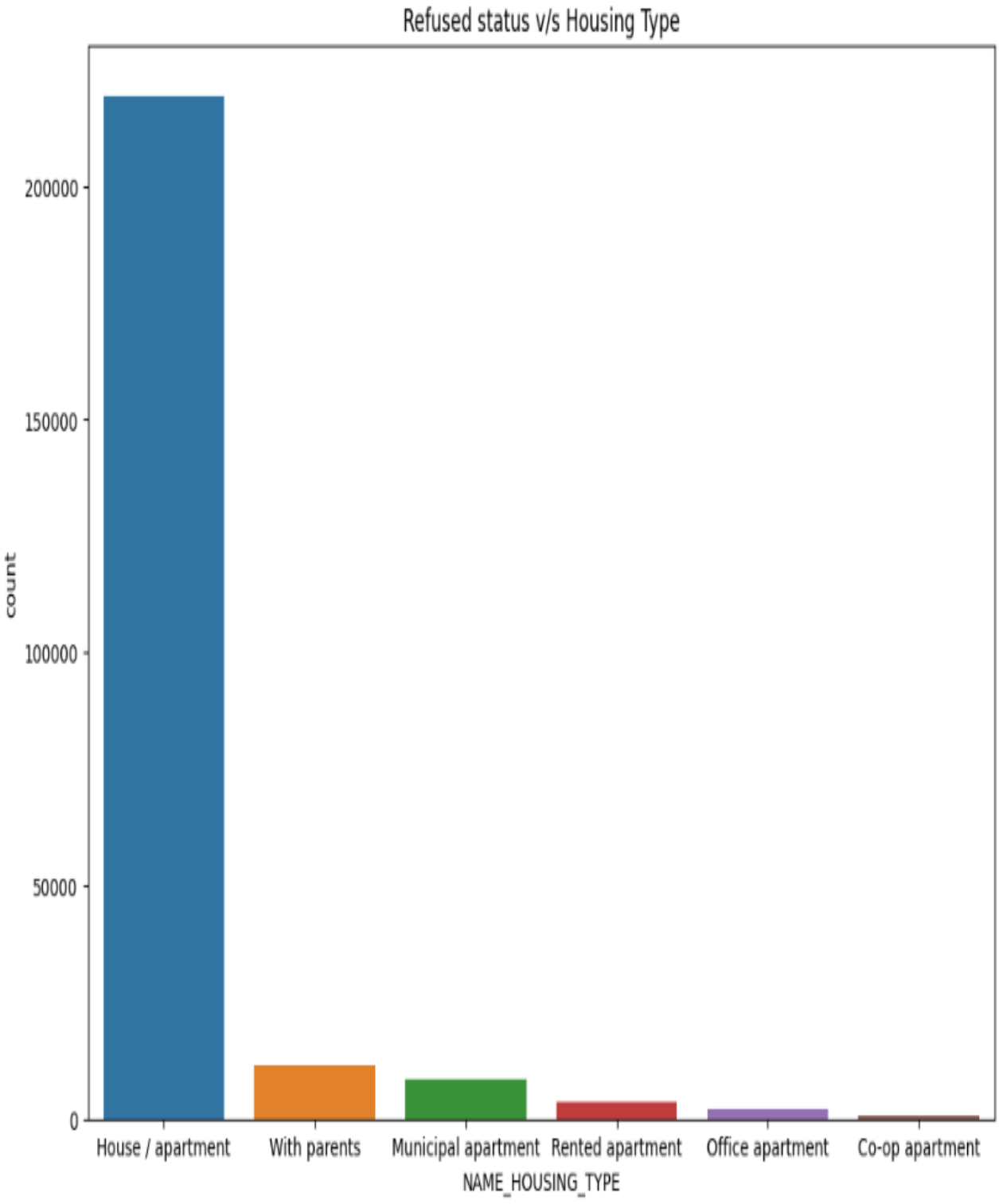

Canceled status v/s No. of Family Members
count
140000
120000
100000
80000
60000
40000
20000
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
12.0
14.0

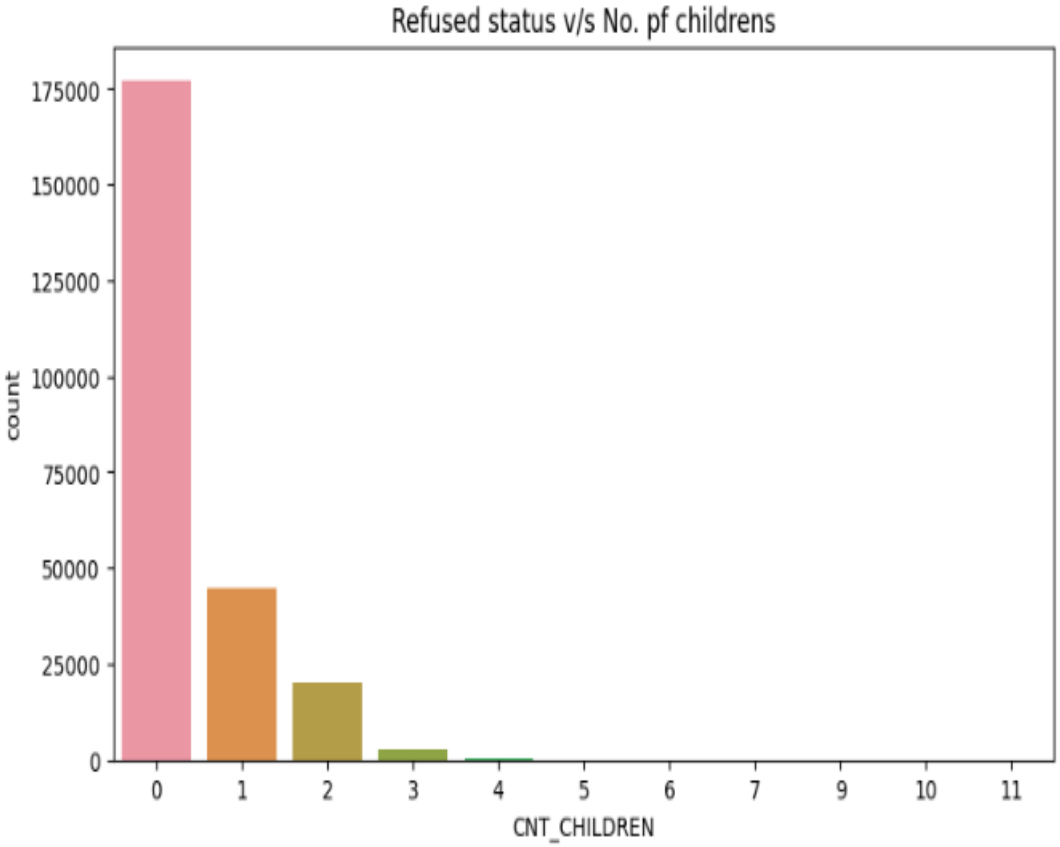


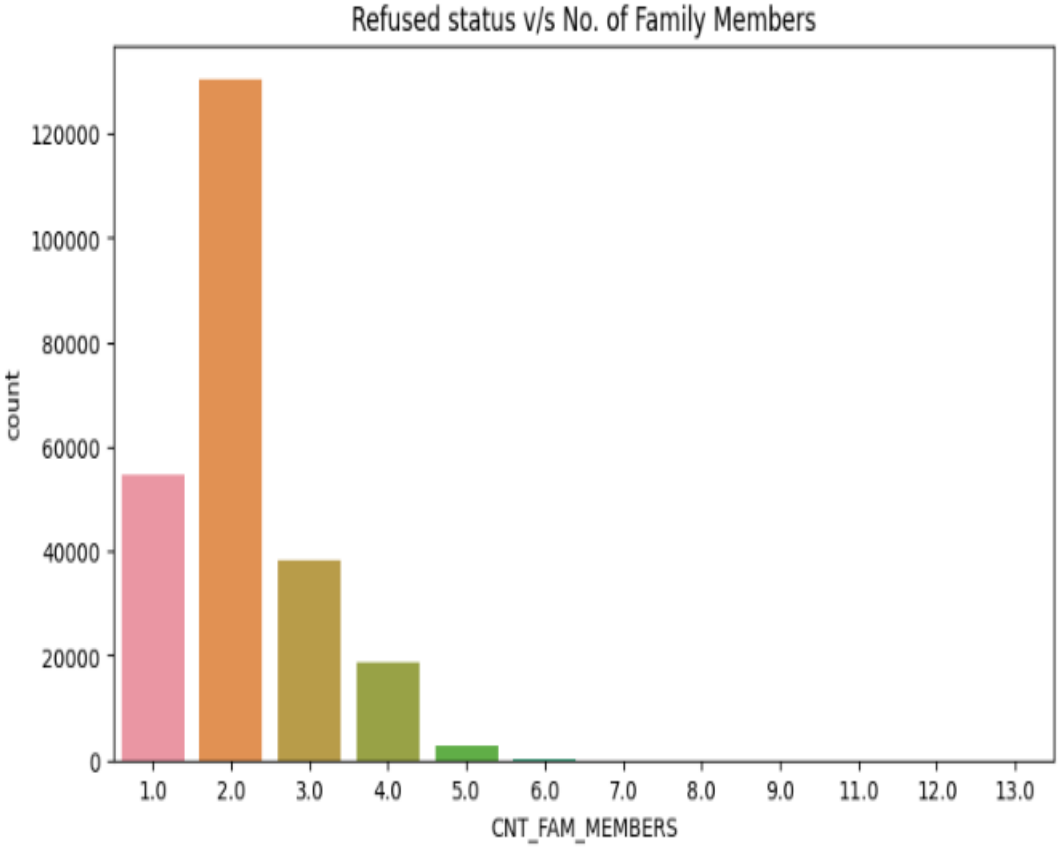
STATUS------REFUSED

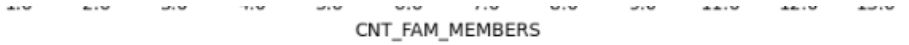


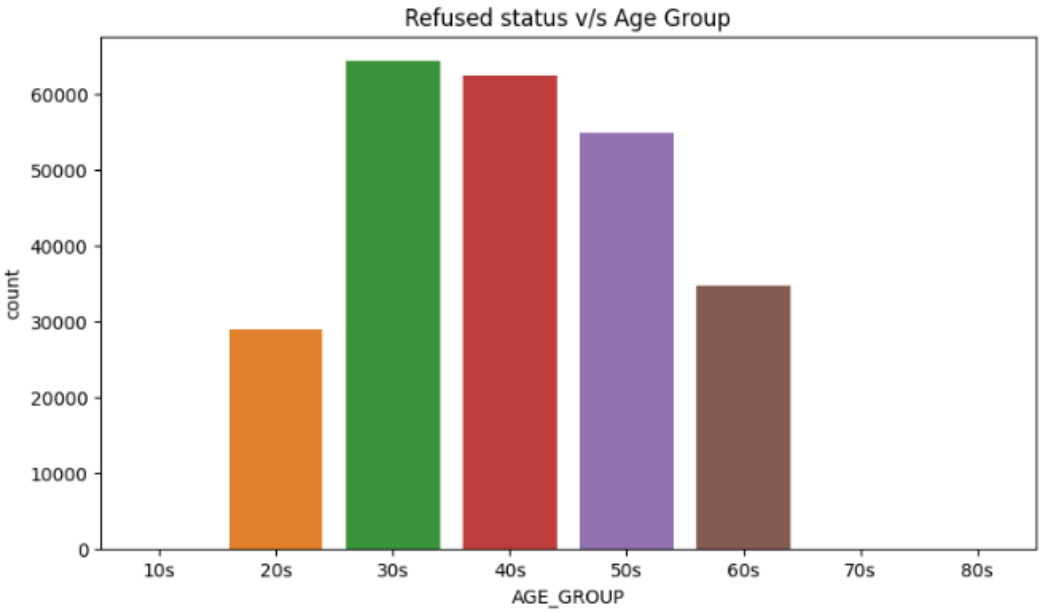



Refused status v/s Housing Type
count
200000
150000
100000
50000
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

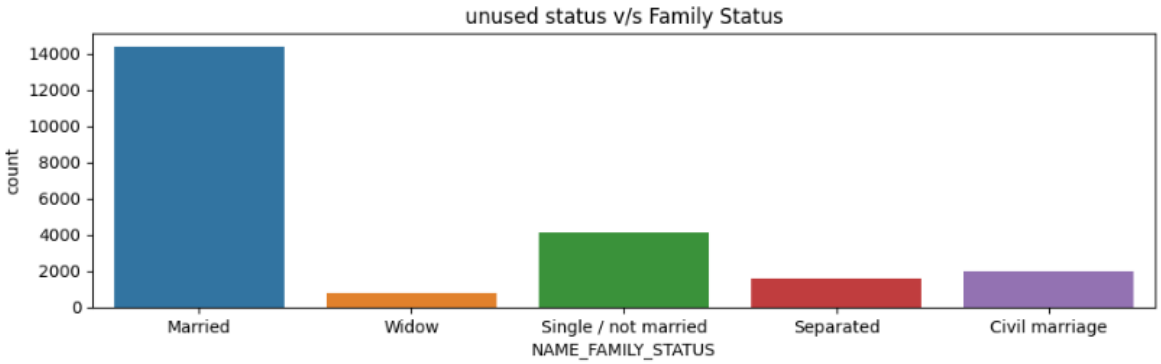

Refused status v/s No. pf childrens
count
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
4
5
6
7
9
10
11
CNT_CHILDREN

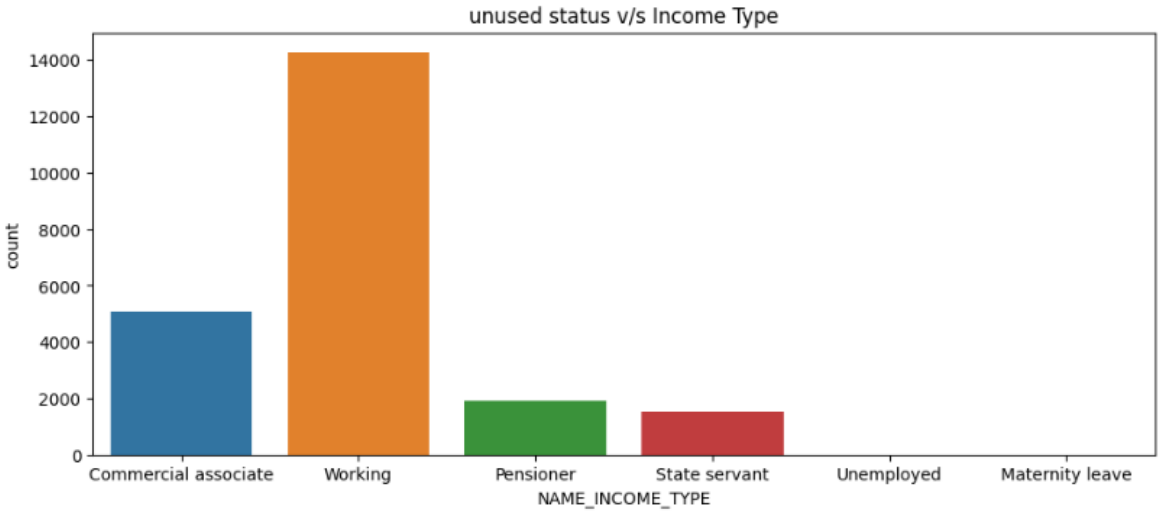

Refused status v/s No. of Family Members
count
120000
100000
80000
60000
40000
20000
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
11.0
12.0
13.0
CNT_FAM_MEMBERS

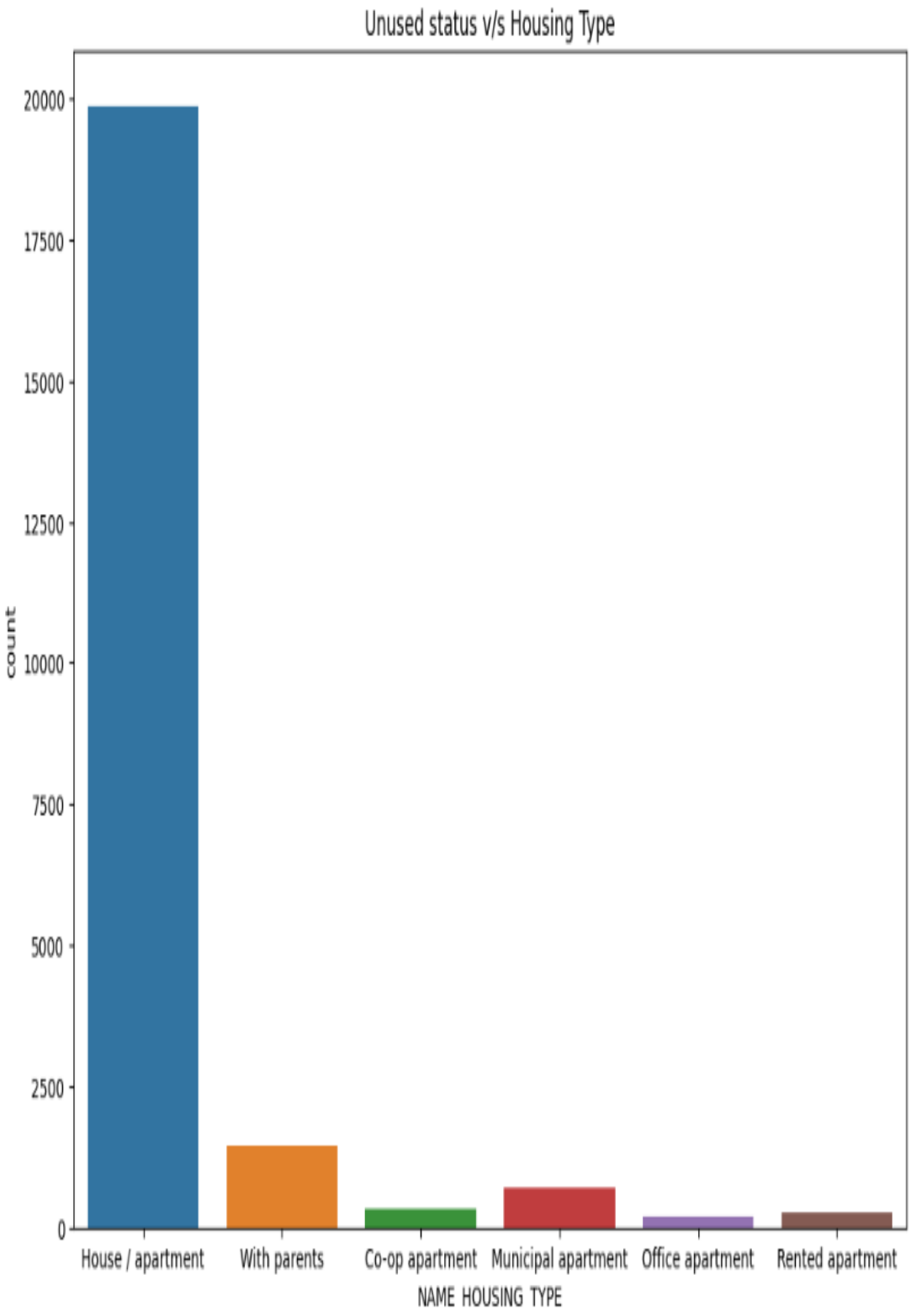


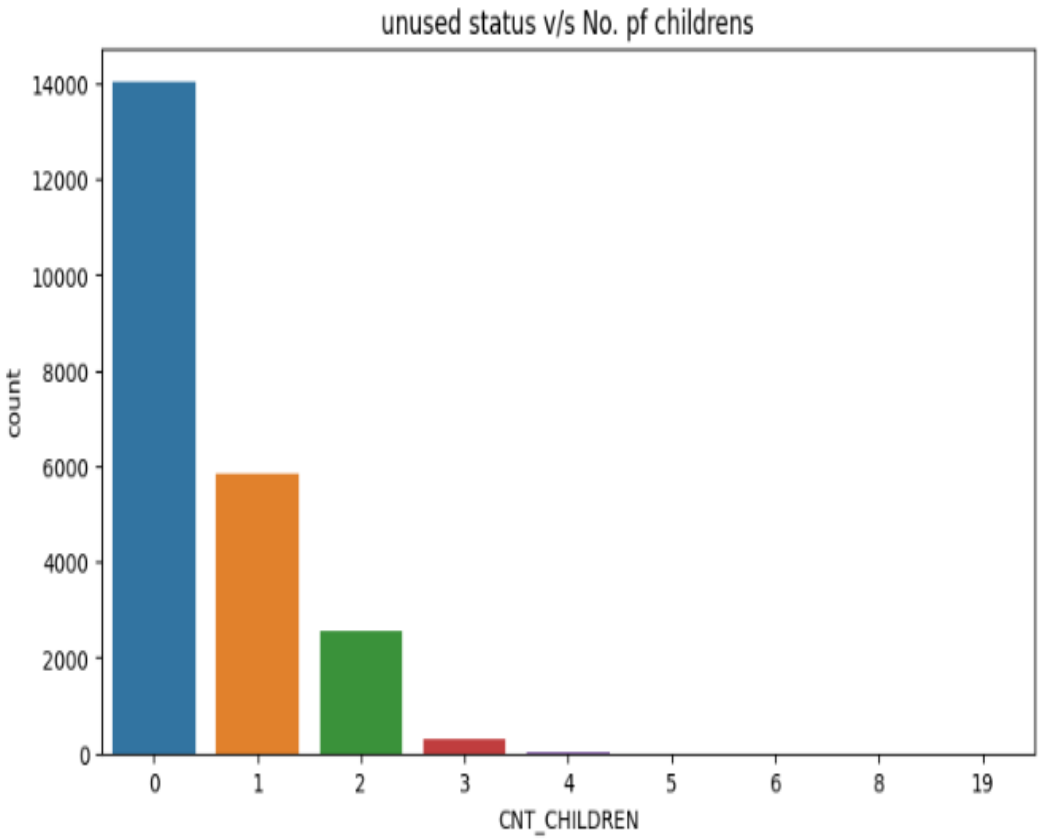


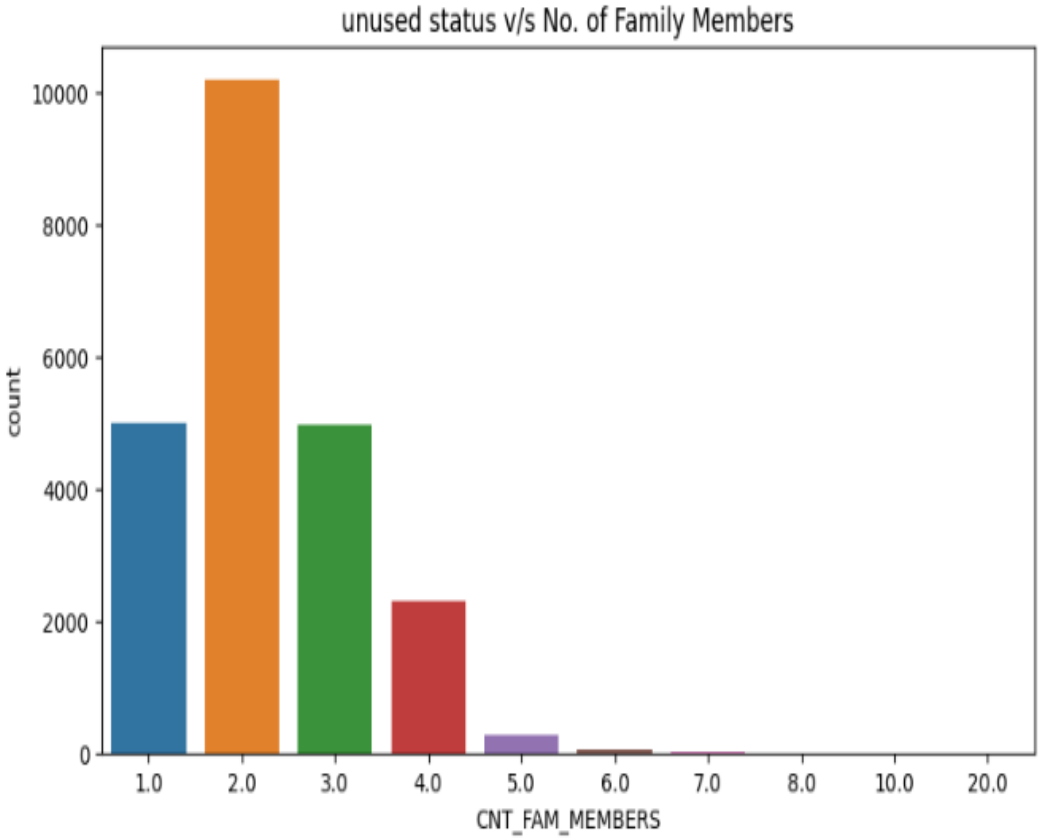
STATUS------UNUSED

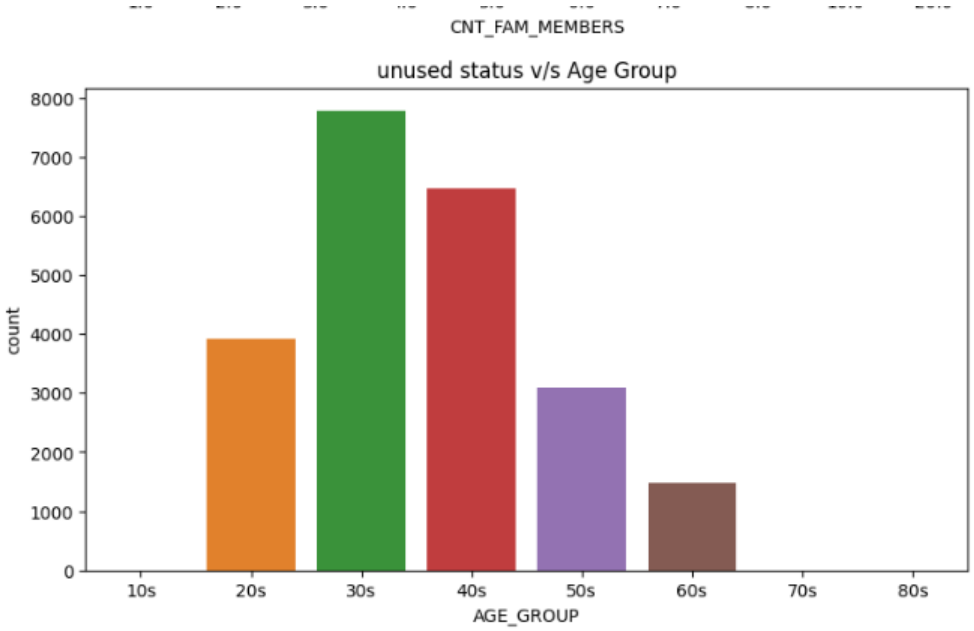





Unused status v/s Housing Type
20000
17500
15000
12500
10000
7500
5000
2500
0
count
House / apartment
With parents
Co-op apartment
Municipal apartment
Office apartment
Rented apartment
NAME_HOUSING_TYPE


unused status v/s No. pf childrens
14000
12000
10000
8000
6000
4000
2000
0
count
0
1
2
3
4
5
6
8
19
CNT_CHILDREN


unused status v/s No. of Family Members
10000
8000
6000
4000
2000
0
count
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
10.0
20.0
CNT_FAM_MEMBERS



Insights from the above analysis for data is as below:

1. Family Status: People who are married are more likely to get loan approved.

2. INCOME Type: People who are working are more likely to get loan approved compared to students who are least likely to get loan approved

3. Housing type: People who own House/apartment are more likely to get loan approved then compared to rented apartments/ co-op apartment types

4. No. of children: People with 0 children are more likely to get loan approved.

5. No. of family members: If the number of people in a family is 2 they are more likely to get loan approved.

6. Age: People with age in between 30 -50 years are more likely to get loan approved compared to the people in 20s and 60s.

Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there.

We explore data sets for target =0 and target =1 to check if top 10 correlated pair of variables are same across both the data sets.

. Top correlate colums are:

- AMT_GOOD_PRICE vs AMT_CREDIT
- AMT_GOOD_PRICE vs AMT_ANNUITY
- AMT_CREDIT_AMT_ANNUITY

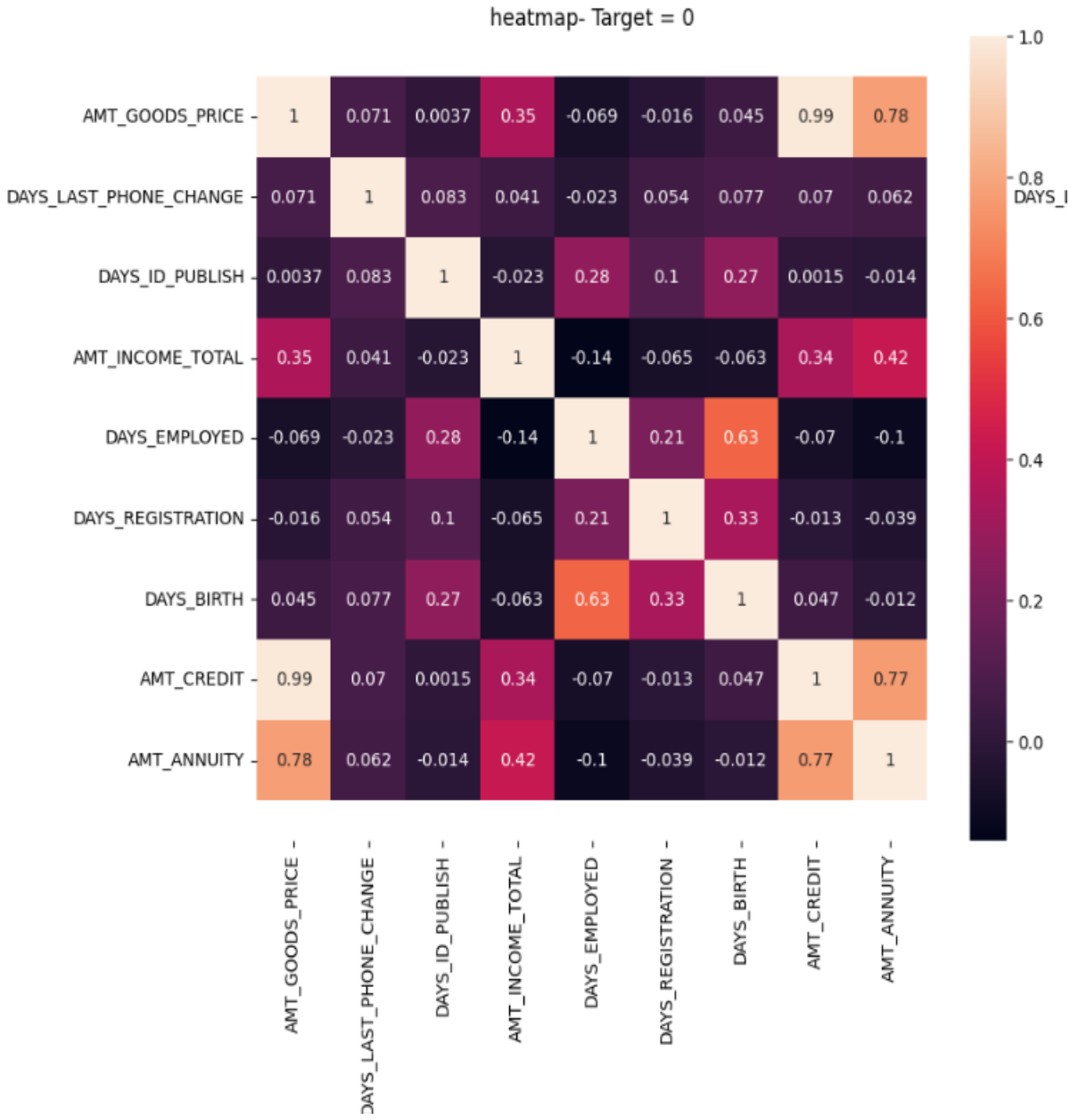


Fig 1: **correlation for variables in target =0**

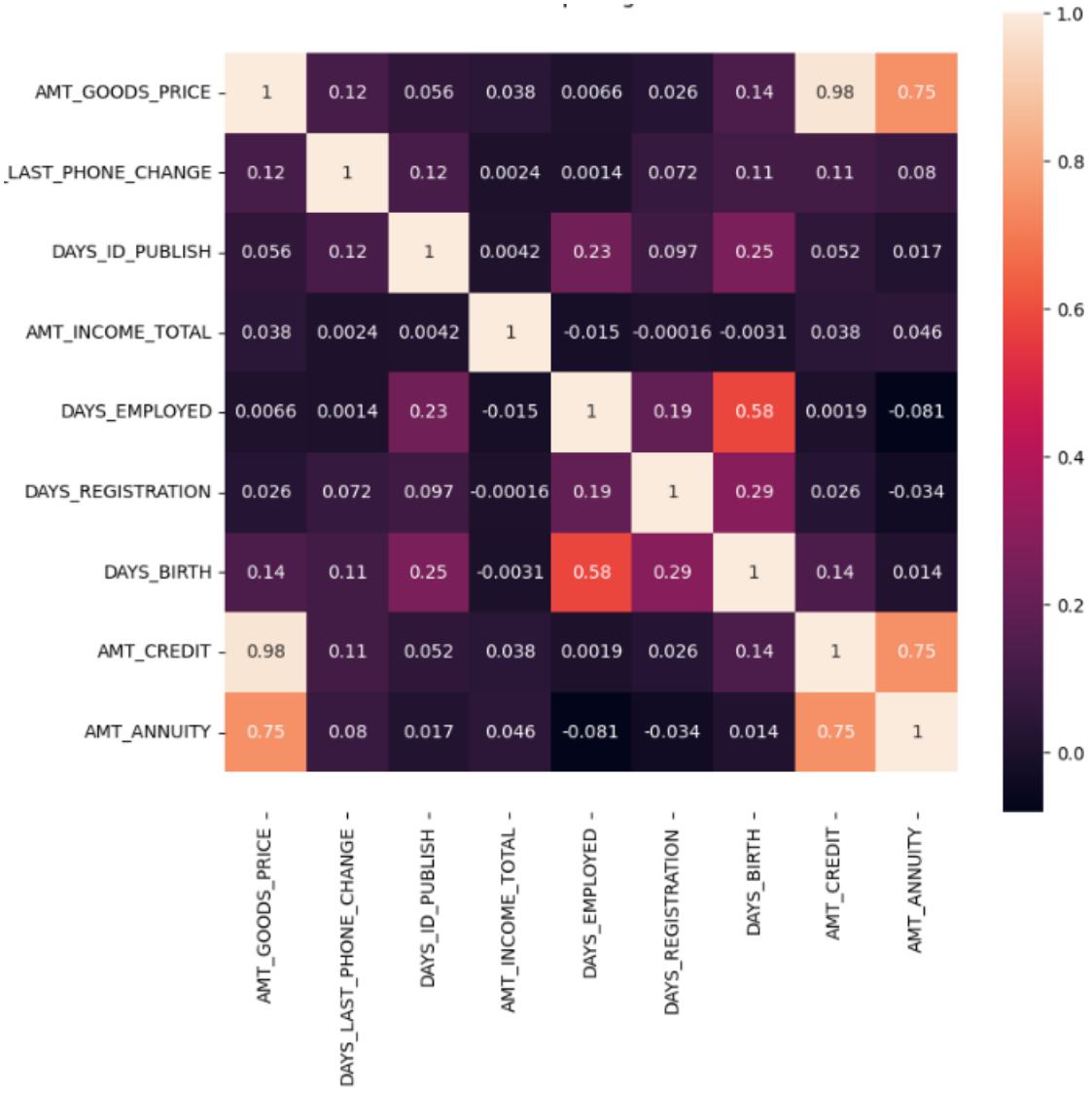


Fig 2: **correlation for variables in target =01**

Though correlation values look similar for both the data sets, we will need to present values in a tubular format to do comparison.

	VAR1	VAR2	Correlation	Correlation_abs
457	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
160	AMT_GOODS_PRICE	AMT_CREDIT	0.99	0.99
377	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95	0.95
314	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88	0.88
484	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86	0.86
161	AMT_GOODS_PRICE	AMT_ANNUITY	0.78	0.78
134	AMT_ANNUITY	AMT_CREDIT	0.77	0.77
242	DAYS_EMPLOYED	DAYS_BIRTH	0.63	0.63
371	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	-0.54	0.54
345	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.54	0.54

Fig 3: **Top 10 correlation variables for target =0**

	VAR1	VAR2	Correlation	Correlation_abs
457	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
160	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
377	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96	0.96
314	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89	0.89
484	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
134	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
161	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
242	DAYS_EMPLOYED	DAYS_BIRTH	0.58	0.58
371	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	-0.45	0.45
345	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.44	0.44

Fig 4: **Top 10 correlation variables for target =1**

8 out of the top 10 correlation variables are common across both the data set.

INSIGHTS:-

- The proportion or percentage of defaulters (target = 1) is approximately 8%, while non-defaulters (target = 0) make up around 92% of the dataset.
- The Bank appears to lend more loans to female clients compared to males. However, the bank can consider targeting more male clients if they meet the credit amount criteria. Male customers have a higher probability of defaulting.
-

- The working class clients have a higher tendency to pay their loans on time, followed by commercial associates.
- Clients with a secondary/higher secondary education level or higher tend to be more prompt in loan repayment. Therefore, the bank may prefer lending loans to clients with such education statuses.
- The age group of 31-40 has the highest count of clients who pay off their loans on time, followed by the age groups of 41-60.
- Married people are safe to target. People having house/apartment are safe to give the loan.
- Clients with lower credit amount ranges have a higher likelihood of paying off their loans on time compared to those with medium or high credit ranges.

RESULTS:-

The project provided valuable insights on how to effectively manage risk when granting loans to clients.

2. The project facilitated the understanding and utilization of various Python libraries such as pandas, matplotlib, and seaborn for data visualization. This enabled the extraction of meaningful information from graphs and charts.
3. The project emphasized the importance of presenting valuable insights and identifying key driving factors from large datasets. It demonstrated techniques to effectively communicate findings and conclusions.
4. An understanding of the correlations between important variables and the ability to present them was gained. This knowledge aids in identifying significant factors that impact loan approvals and risk assessment.
5. The project covered various data preprocessing techniques, including handling null values, imputing missing data, addressing data imbalance, and detecting outliers. Additionally, it involved performing univariate and bivariate analysis to explore relationships and patterns within the data.
6. The project provided practical experience in utilizing Exploratory Data Analysis (EDA) techniques in real business scenarios. It highlighted the relevance and applicability of EDA in understanding data and making informed decisions.

Overall, the project contributed to a deeper understanding of risk management in the context of loan applications, as well as the application of various data analysis techniques and tools in a practical business setting.

GOOGLE COLAB LINK----------
https://colab.research.google.com/drive/11slh5hNJeaR0Lkf3of2igaCtAXVpKLAD?usp=sharing